Adapting Quantitative Protein and Phosphorylation Analyses to a Proteome-Wide Scale

Adapting Quantitative Protein and Phosphorylation Analyses to a Proteome-Wide Scale

A dissertation presented

by

Joshua Terrence Wilson Grady

to

The division of medical sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Cell Biology

Harvard University

Cambridge, Massachusetts

April, 2013

Dissertation Advisor: Professor Steven P. Gygi Author: Joshua Terrence Wilson Grady

Adapting Quantitative Protein and Phosphorylation Analyses to a Proteome-Wide Scale

Abstract

Liquid chromatography coupled tandem mass spectrometry (LC-MS/MS) has become the preferred method for large-scale peptide and phosphopeptide identification and quantification. The dominance of LC-MS/MS is the result of improved chromatographic, mass spectrometry and bioinformatic technologies. The applications of these technological improvements drive biological innovation by expanding the realm of possible experimentation, facilitating the creation and evaluation of novel hypotheses. Such improvements are the focus of this dissertation. New technologies are presented and their proteome wide applications in biological systems are demonstrated.

A comparison of common phosphopeptide enrichment methods is presented in chapter two, which demonstrates that a combination of methods provides non-overlapping data sets. This comparison was performed in mitotically arrested fission yeast, a previously unstudied system by phosphoproteomic methods. This chapter remarks upon phosphorylation site conservation between lower and higher eukaryotes, as a means of predicting potentially relevant phosphorylation events in mammals.

A new protocol for tissue based peptide quantification is presented in chapter three. The large-scale application of this method is detailed in a system of mouse liver phosphorylation, between fasted and re-fed states. The effect of peptide and protein level false discovery rates on the accuracy of phosphorylation site quantification is highlighted. This method is a cost-effective alternative to available techniques, such as metabolic labeling, and expands the application of proteomics to include larger animals.

Finally, an in depth analysis of quantitative LC-MS/MS based multiplexing is the subject of the last chapter. New techniques for peptide pre-fractionation and ion quantification are discussed, which improve proteome coverage and quantitative accuracy. This proteome-wide multiplexing is applied to an analysis of the budding yeast environmental stress response. Applicable methods of data processing and a means of obtaining biologically relevant information out of multidimensional proteomic data sets are discussed. In all chapters, the data presented represent the largest analyses of their kind. This dissertation provides a solid guide for future proteome-wide studies, focused on the identification and quantification of peptides and their posttranslational modifications.

Table of Contents

List of Figures

Chapter 1: Introduction

Chapter 2: Phosphoproteome Analysis of Fission Yeast

Chapter 3: Quantitative Comparison of the Fasted and Re-fed Mouse Liver Phosphoproteomes Using Lower pH Reductive Dimethylation

Chapter 4: Proteome-Wide Applications of Quantitative Multiplexing in the Yeast Stress Response

List of Tables

Acknowledgments

My endeavors in graduate school, though they required much self-motivation and work by my hands, could not have been completed without the support of my colleagues and close companions. My success is a direct reflection of their efforts. I am indebted to many people within the lab and the Harvard community in general. First and foremost, Dr. Steven Gygi has provided an academic environment which is unparalleled. I cannot think of another person in the field of mass spectrometry based proteomics who has the combination of vision, enthusiasm, and skill which has made his lab so successful. Dr. Gygi's guidance has both permitted my comprehensive education and encouraged my love of scientific exploration. Though Dr. Gygi's influence has been great, his ability to draw high quality scientists to his lab is equally as important.

I owe much of the success in my graduate education to three people from Dr. Gygi's in particular, Drs. Judit Villén, Wilhelm Haas, and Edward Huttlin. Dr. Villén and I worked closely at the beginning of my tenure in the Dr. Gygi's lab, and she gave me a solid foundation for large scale phosphopeptide analysis. Dr. Haas has been a second mentor to me and has greatly contributed to my understanding of LC-MS technologies. In addition, Drs. Haas and Villén were primarily responsible for instrument maintenance during the majority of my graduate career. Dr. Huttlin is particularly skilled in bioinformatics and statistics, and as such, has opened my eyes to many methods for large data processing and analysis. Other members of the Gygi lab have also left their mark on my career. In particular I would like to acknowledge Drs. David Nusinow, Graeme McAlister, Woong Kim, Noah Dephoure, Robert Everley, Lily Ting, Martin Wühr and Mark Jedrychowski for helpful discussions, guidance, reagents, instrumentation assistance, and bioinformatic support. I am also indebted to the "GFY Development Team," which provided the bioinformatics platform for the analysis of LC-MS data.

This team includes, but is not limited to, Dr. Joshua Elias, Dr. Corey Bakalarski, Dr. Julian Mintseris, Dr. Sean Beausoleil, Deepak Kolippakkam, Ramin Rad, Trent Ostler, Dr. Ed Huttlin and Dr. David Nusinow.

I would also like to thank my dissertation advisory committee - Drs. John Blenis, Randy King and Wade Harper- for their helpful advice. Their guidance allowed me to focus on the aspects of my projects which were most likely to succeed, thereby avoiding pitfalls which may have unnecessarily prolonged my graduate studies. Their much needed criticism has taught me to evaluate my intended research goals, and create an efficient plan to reach these goals.

I am also very grateful for the system of public education that we have in the United States of America, and the phenomenal educators who comprise this system. Prior to attending Harvard Medical School, my education exclusively occurred within public institutions. At these institutions I received the highest caliber of education, and my graduate education would not have been possible without this infrastructure. In addition, the majority of my graduate funding was through the National Institute of Health, reiterating the function of the government in my education. The continued need for these opportunities is crucial for the education of future generations.

Finally, the people who deserve the most credit for my success have always been, and always will be, my family. My parents, Martin and Christine Wilson-Grady, provided a supportive environment, which encouraged me to think critically about the world around me. From them, I developed the basic skills upon which my scientific education was based. As part of my support structure, I have been fortunate to always have a best friend in my brother Nathaniel Wilson-Grady; he has always been there for me when I need him. Though my education was greatly supported by all members of my family, above all else, it is my wife, Dr. Nicole Grady, who has made the most impact in my life. She is a brilliant woman, a phenomenal wife, a caring mother, and generally the most loving, supportive person I know. I cannot fathom completing such a long journey without her unwavering support. Particularly over the last few months, after the birth of our amazing daughter, Abigail, Nicole made my work possible by

taking on the majority of responsibility. I am eternally indebted to her. She inspires me to be the best man I can be in all facets of my life.

Chapter 1

Introduction

Proteomics: a 21st Century Technology

The last three decades of the twentieth century saw the rise and maturation of the genomic sciences. What began as laborious and costly effort for even simple genomic analyses[1, 2], flourished into a robust strategy, applicable in complex organisms. Indeed the turn of the century witnessed the completion of the human genome[3, 4], a milestone with far reaching potential. In addition to gene sequencing technology, the development of robust microarray strategies has produced libraries of gene expression data, encompassing various model organisms and human diseases (e.g. ref 5). The interpretation of genetic sequence information and gene expression data, however, is not always readily comprehensible. Proteins are primarily responsible for the direct actions within a cell, and often gene and protein expression patterns display limited correlation[6]. Furthermore, posttranslational events, particularly phosphorylation, may regulate the functions of these proteins, complicating the biology at hand. It is evident that robust strategies for the identification and quantification of proteins and their posttranslational modifications are of extreme use for the analysis of biological systems. In an analogous manner to genomics, the field of proteomics has developed over the last two decades, beginning as a low throughput qualitative endeavor, and maturing to a large-scale quantitative technology.

The common theme between genomics and proteomics is that technology drives biology. With the introduction of new technology (e.g. multiplexed DNA sequenceing[7] and absolute protein quantification methods[8]), novel experiments become possible which can answer fundamentally unique questions. In addition, the results of these experiments often generate a wealth of new hypotheses in a

manner which traditional small-scale experiments cannot. In this dissertation, the development, evaluation, and the biological application of proteomic methods and technologies are discussed, with an emphasis on posttranslational modification and peptide quantification. The foundation of each chapter is a solid understanding of shotgun sequencing by LC-MS/MS and the incorporation of stable isotopes for mass-based quantification. An overview of proteomic technologies and a presentation of these topics are discussed in depth.

Common Analytical Techniques in the Field of Proteomics

Although peptides have been characterized for many decades by using analytical tools such as amino acid composition analysis or Edman degradation[9], sensitive and high-throughput technologies were slower to develop. Modern protein analysis utilizes several technologies, of which various strengths and weaknesses exist: these include one- and two-dimensional gel electrophoresis, western blotting, and protein arrays (summarized in Table 1.1). Often these technologies are limited in their sensitivity, specificity and depth of analysis, or by other means, such as the antibody requirements of western blotting and protein arrays. These drawbacks hinder their application on a proteome wide scale. HPLC coupled tandem mass spectrometry (LC- MS/MS) offers the most promise for overcoming these limitations, and is an intense are of research[10]. The coupling of LC and MS components[11] gave rise to the 'shotgun sequencing' technique, which is now the preferred method for peptide analysis.

Table 1.1. Comparison of common methods for peptide and protein analysis

Method	Identified Molecule(s)	Throughput	Detection Limit	Advantages	Disadvantages
Amino acid analysis	Amino acid composition	2 days/analysis	Picomoles	Very accurate	Slow, requires pure peptide samples
Edman degradation	Peptide sequence	1 hr/residue	Picomoles	Generates *De novo* sequence information	Slow, requires pure peptides,low fidelity
1D and 2D SDS-PAGE	Protein migration patterns	1-2 hours	Hundred of picograms	Low cost, ubiquitously applicable, proven technology	Difficult to detect low abundance proteins, proteins may co-migrate
Western blotting	Epitope	4-6 hours	Femtograms	Sensitive, ubiquitously applicable, proven technology	Requires antibodies; slow to develop if unavailable, semi-quantitative
Protein arrays	Epitope	Thousand of proteins/analysis	Low picograms	High-throughput, automated, sensitive	Difficult to construct array, developing technology, requires extensive validation
LC-MS	Peptide sequence	Thousand of peptides/hour	Femtomoles	High-throughput, sensitive, specific, quantitative	Sample complexity affects analysis and reproducibility, requires extensive bioinformatics

Overview of LC-MS/MS, Sample Preparation and Proteome Complexity Considerations

Sequencing peptides by LC-MS/MS is multistep, yet linear process. Peptides are separated by online reverse phased liquid chromatography and ionized as they elute off the column. The mass to charge ratio (m/z) of ionized peptides is detected in an MS^1 scan, ions of interest from an MS^1 scan are sequentially selected for fragmentation[12], and fragment ions are reanalyzed in an MS^2 scan. The resulting MS^2 spectra are searched by means of various computer algorithms to obtain peptide sequence information (the typical workflow of shotgun sequencing is displayed in Figure 1.1). Often these algorithms identify peptides by matching the observed MS/MS spectra to lists of potential theoretical spectra[13] (Figure 1.1, B). This strategy has been successfully applied in numerous analyses, with the caveat that a peptide can only be matched if its sequence is contained in the database. A peptide which is not contained in the database will be incorrectly matched, increasing the false discovery rate of the analysis. Much effort has been put toward understanding and controlling false discovery rates[14-16]. As proteome size increases, and as common peptide modifications are included in the search criteria (e.g. phosphorylation), the complexity of these searches increases by several orders of magnitude. Mass accuracy has greatly aided in peptide identification in such cases. All of these considerations are elaborated upon later.

In the vast majority of proteomic analyses[10-12, 17-19], protein lysates are digested with endopeptidases prior LC-MS/MS. In general, peptide analysis is preferred, as peptides tend to be shorter and less charged than their protein counterparts; large and highly charged molecules are not amenable to mass spectrometry without the application of specialized techniques (e.g. the 'top down appoach'[20]). Trypsin (cleaves after K and R residues) and Lys-C (cleaves only after K residues) are two common proteolytic enzymes used for digestion. These enzymes generally liberate peptides between 10-20 amino acids, carrying 2-4 charges in the gas phase, making them well suited for mass spectrometric applications.

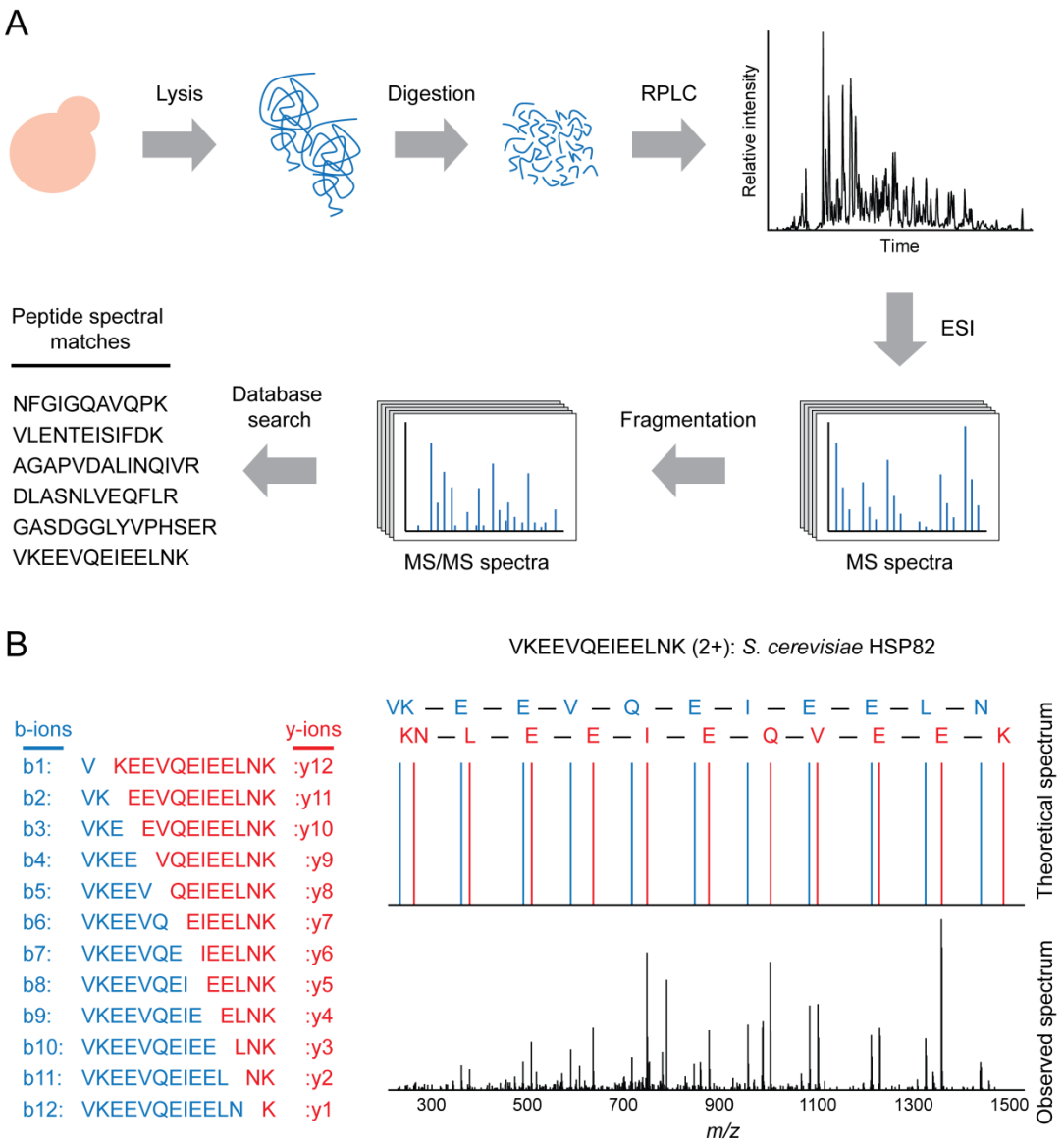


Figure 1.1. Workflow of a typical shotgun sequencing experiment by LC-MS. (A) Data acquisition strategy. The first step in the analysis of peptides by shotgun sequencing is to obtain a whole cell lysate from the cell or tissue of interest. In this example, budding yeast are analyzed. Protein lysates are digested by endopeptidases to produce sequence specific fragments. Often either trypsin (cleaves C-terminal to K and R) or Lys-C (cleaves C-terminal to K only) enzymes are used for digestion. The digested peptides are separated by online HPLC, using reverse phased chromatography in a glass capillary column. As peptides elute, they are ionized into the mass spectrometer through electrospray ionization (ESI). These ionized peptides are analyzed within a defined *m/z* range to produce an MS^1 spectrum. Ions of interest (usually the top 10-20 most abundant ions) from an MS^1 spectrum are sequentially isolated and fragmented (e.g. by CID). These fragment ions are reanalyzed, producing an MS^2 spectrum. This MS^2 spectrum is searched against a protein database to obtain peptide sequence information using one of several available search algorithms. (B) Spectral matching of observed and theoretical spectra by the SEQUEST algorithm. CID generates MS^2 spectra which contain b- and y-type fragment ions, for the N- and C-terminal fragments respectively. These ions are matched to the theoretical b- and y-ions of a potential spectrum. The degree to which the theoretical and observed spectra correlate (SEQUEST XCorr) is a measure of how certain one can be of a peptide's identification.

Peptides from whole cell lysates may be either analyzed directly, or additional fractionation techniques may be used prior to LC-MS/MS (e.g. strong cation exchange chromatography). Alternatively, protein fraction by SDS-PAGE, for example, may be used prior to digestion. Although the online liquid chromatography component effectively separates peptides from one another (increasing analytical depth) while concentrating each peptide in the capillary column (increasing analytical sensitivity), often the complexity in peptide concentration range of an unfractionated proteome digest is large[21]. The result of this great complexity is the co-elution of peptides and stochastic identification, resulting in fewer identified peptides. This phenomenon is probably the largest drawback to shotgun sequencing. The function of pre-fractionation techniques is a reduction in sample complexity to reduce stochastic sampling. Faster scanning and more accurate mass spectrometers have also been demonstrated to lessen the impact of sample complexity, as has the use of dynamic exclusion algorithms, which actively avoid re-isolating the same ions within a defined period of time. Each sub-process of shotgun sequencing technology is an area of intense research.

Peptide Ionization

Although many ionization techniques exist in the field of mass spectrometry, including matrix assisted laser deportation ionization (MALDI)[22] and electrospray ionization (ESI)[23], ESI is the preferred method for peptide analysis by LC-MS. As peptides elute off the reverse phased column, they are passed through an electric potential, which ionizes peptides (bestowing a positive charge upon them) and vaporizes the mobile phase, creating microscopic droplets. The use of ESI is preferred for several reason: first, multiple charge states are formed, which extends the possible mass to charge (*m/z*) range of the analysis; MALDI in contrast tends to primarily form singly charged ions, limiting its usefulness for complex mixtures. In addition ESI generally avoids molecular fragmentation during ionization. Finally,

only one analyte is contained within each ionized droplet on average, facilitating their efficient transfer to the gas phase for MS analysis.

Modern LC-MS/MS analysis employs a 'nanospray' technique, which involves the use of thin columns (100 µm inner diameter or smaller), small stationary phase resins (particle size of 3 µm), high pressure and low flow rates (~300-500 nL/min) [24, 25] for analysis. This technique demonstrates improved chromatography over other methods, which leads to greater peptide resolution and sensitivity[25]. The discussed methods are coupled to automatic sampling technology, which automates the procedure and increases the analytical throughput.

Mass Analysis with Parts per Million (PPM) Accuracy

One of the greatest technological achievements in the field of mass spectrometry based proteomics, during the first decade of the 21st century, was the implementation of instrumentation which permitted the determination a peptide ion's *m/z* with extremely high accuracy and resolution. High resolution instrumentation is useful for accurately visualizing peptide isotopes, which contain very similar differences in their mass to charge ratios (*m/z*). For example, with the commercial availability of Fourier-transform ion cyclotron resonance (FT-ICR) mass spectrometers, peptides could routinely be analyzed with parts per million (PPM) accuracy[26]. Using FT-ICR, determination of an ion's *m/z* is based on its cyclotron frequency in a magnetic field. A packet of ions is excited using a specific radio frequency voltage (corresponding to the cyclotron resonance of those ions), which gradually increases the orbital radius of the ion packet. Once this radius is suitably large, the ions induce a current pulse in detector electrodes as they orbit, which can be plotted over time. This time domain is Fourier transformed to yield the mass of these ions. Such a process occurs in a multiplexed fashion, where many packets are simultaneously excited and detected, which creates the full mass spectrum by the combination of Fourier transformed signals[27]. Such accuracy obtained through FT-MS is ubiquitously helpful for peptide

identification, the identification of posttranslational modifications, and is required for peptide quantification. Recently, the use of an Orbitrap for high accuracy mass analysis[28] has become the preferred method. Although both the Orbitrap and the FT-ICR are capable of high accuracy FT-MS, the Orbitrap is smaller, less expensive, and requires less maintenance; the FT-ICR requires the use of a liquid helium cooled superconducting magnet to generate the magnet field, whereas the Orbitrap utilizes electrostatic fields for trapping ions.

High accuracy mass measurements substantially improve the identification of peptides by LC-MS/MS. The most obvious advantage with high mass accuracy is the limitation of theoretical peptide sequences which may match an MS^2 spectrum, as smaller mass tolerances in the search parameters may be used. The result of limiting the number of potential spectra is a reduction in computational time for search algorithms. In addition, the reduction of theoretical candidates is an effective means of controlling the false discovery rate (discussed below) of a proteomic analysis, as it becomes increasingly likely that false peptide assignments will match a given MS/MS spectrum as the search space is increased[15]. Thus mass accuracy leads to higher sensitivity in the assignment of MS^2 spectra, as more of the spectra can be successfully matched to true positive sequences. Related to this fact, high accuracy analysis resolves peptide isotopic envelopes (due to naturally occurring ^{13}C in proteins), allowing for the unambiguous assignment of charge state based on the *m/z* difference between successive isotopic peaks. As peptides are measured as a function of their mass to charge ratio, accurate identification of the charge state allows one to back calculate the peptide mass with precision. Without mass accuracy, peptide spectra are searched in multiple charge states, which complicates the search and may affect the false discovery rate[15]. These observations are particularly true for the analysis of posttranslational modifications, as the inclusion of common modifications such as phosphorylation, oxidation and ubiquitination, greatly increases the number of potential theoretical spectra by several orders of magnitude. High mass accuracy also is extremely beneficial for peptide quantification using stable

isotopes (discussed in detail later). Accurate mass readings allow one to extract ion intensity information from several successive MS^1 scans, generating extracted ion chromatograms. In this manner an MS^2 is not required for each MS^1 ion of interest to quantify a peptide[29]. Smaller mass windows also limit the amount of observed noise, increasing peptide signal to noise measurements and their quantitative accuracy.

High accuracy analysis (e.g. via the Orbitrap) is now routinely combined with fast scanning mass spectrometers (quadrupole liner ion traps, e.g. LTQ) in hybrid instrumentation (LTQ-Orbitrap). In these instruments, a large mass range of ions is first trapped in the LTQ and a full MS (MS^1) scan of these ions is read out in the Orbitrap at high resolution. Subsequently, ions of interest are re-acquired and isolated in the LTQ (~2 *m/z* isolation windows around the ion of interest), and fragmented (methods discussed below). These product ions are rapidly analyzed in the linear ion trap to obtain MS^2 spectra for database searching. MS^2 product ions tend to be low abundance species and benefit from the sensitivity imparted by the linear ion trap. The result of this instrument hybridization is a combination of high accuracy MS^1 with fast and sensitive acquisition of MS^2 spectra. This compromise is an ideal solution for the analysis of complex mixtures, and has led to deeper proteome analyses. One current implementation of hybrid instrumentation is the Orbitrap Velos; in this instrument, two linear ion traps are used in series, one for fragmentation (high pressure trap) and one for product ion analysis (low pressure trap)[30]. MS^1 analysis still occurs in the Orbitrap. This instrument provides a substantial increase in analytical depth compared to its predecessor.

It is additionally possible to analyze MS^2 fragment ions with high accuracy in the Orbitrap (parts per million); in contrast, ion trap MS^2 spectra are lower accuracy (parts per thousand). High accuracy MS^2 spectra contain several benefits: The combination of high accuracy MS^1 and the identification of even a few high accuracy MS^2 ions may drastically limit the number of theoretical peptides which could potentially match an MS^2 spectrum; in many cases only one or a few peptides may be possible. Search

algorithms are further aided by such an analytical scheme, and peptides often score better with these algorithms. In addition, high accuracy MS^2 spectra improve the performance of fragment ion based peptide quantification techniques (e.g. TMT, discussed later, and the subject of chapter 4). Finally, the combination of high accuracy MS^1 and MS^2 spectra allows one to perform unconventional searches which are achieved using a large (100 Da) parent ion (MS^1 ions) tolerance. In these spectra (used briefly in chapter 3) it is possible to visualize peptide modifications which have not been previously reported (e.g. reaction side products) or simply not accounted for in the search parameters (e.g. deamination of asparagine and glutamine). The mass of the modification is observed as the mass of the observed peptide minus mass of the theoretical peptide sequence (mass error, can be negative). Despite the large mass error, peptides are successfully identified as results of the high accuracy MS^2 ions, an outcome which is limited with ion trap MS^2.

Most applications of high accuracy MS^2 have been limited by the decreased sensitivity and increased analytical time required to collect such spectra. That being said, the recent implementation of the Q-Exactive mass spectrometer (a quadrupole mass filter coupled to an Orbitrap analyzer[31]) has greatly increased the application of high accuracy MS^2 analysis, due to decreases in the MS duty cycle. Ion sensitivity, however, is still an issue with this mass spectrometer, thus generally low intensity samples (e.g. phosphorylation) are not analyzed by this means.

Fragmentation Options

Regardless of what type of mass analyzer is used, all shotgun sequencing requires peptide fragmentation by tandem mass spectrometry (MS/MS), to obtain sequence information. As mentioned, MS/MS occurs by first selecting and isolation an ion of interest from the MS^1 spectrum. The isolated ions are then fragmented by one of many available methods, producing characteristic fragment ions (fragmentation along the peptide backbone). The resulting fragment ions are then analyzed to obtain an

MS^2 spectrum used for database searching and peptide identification. Generally the top 10 to 20 precursor ions in an MS^1 spectrum are sequentially selected (in decreasing order of abundance) for MS^2 analysis, depending on the instrument used and the goal of the analysis[12, 32]. The following descriptions of fragmentation methods are as they pertain to MS/MS analysis on a LTQ-Orbitrap platform. These methods are applicable to other instrumentation platforms, though their implementations are slightly different. Three common fragmentation methods are collision induced dissociation (CID), electron transfer dissociation (ETD), and higher energy collisional dissociation (HCD). These methods are generally complementary with respect to one another, with each demonstrating advantages under certain conditions.

CID is perhaps the most widely used peptide fragmentation method in mass spectrometry. In CID, peptide ions are activated (within a defined *m/z* window) and collided with an inert gas (usually helium or nitrogen). The result of these collisions is fragmentation along the peptide backbone. The collisions tend to produce only one fragmentation event per peptide, as the *m/z* of peptide fragments generally fall outside of the activation window for CID, and they are therefore not reactivated for further collision. As a result many b- and y-type ions are produced (N- and C-terminal fragments). In the discussed hybrid instruments, CID occurs in the linear ion trap, and fragment ions can be analyzed either in the linear ion trap (fast/sensitive but lower mass accuracy) or the Orbitrap (slow/less sensitive but high mass accuracy). CID can also be performed using other common instruments, such as triple quadrupole mass spectrometers. The advantages of each MS^2 analysis method are discussed above.

ETD fragments ions through the transfer of radical anions to protonated peptides, which cause fragmentation at the Cα-N bond[33]. As such, analogous c- and z-type ions are generated, as opposed to the b- and y-type ions generated by CID. ETD has been presented as a promising fragmentation method, particularly for the analysis of phosphopeptides. Labile phosphate groups are often liberated during CID (neutral loss), complicating the localization of a phosphate group as sequence-specific b- and y-ions in

these MS^2 spectra often of low intensity. In contrast neutral loss in generally not observed with ETD. At least in our lab, however, we find that CID generally outperforms ETD, despite factors such as neutral loss, due to the slower speed of ETD analysis. As such it is not generally discussed further.

HCD on the other hand is an alternative to CID which is now widely used for proteomics studies[30, 31], particularly for MS^2 based peptide quantification[34]. HCD is a type of CID, in that peptide ions are collided with inert gas molecules to produce fragment ions (b- and y-type). In contrast to ion trap CID, HCD utilizes higher energy beam-type fragmentation[35]. These high energy dissociations enable a wider range of fragmentation pathways, where multiple fragmentation events per peptide ion are possible. A drawback of ion trap CID is that a low mass cutoff exists (~30% of the low *m/z* range of the isolation ion is lost), hindering the analysis of smaller fragments. HCD is not restricted by this property and is useful for the analysis the low mass reporter ions in MS^2 based quantification (discussed later). In most instruments, HCD is generally performed in a dedicated collision cell and the ions are analyzed with high mass accuracy in the Orbitrap.

Database Searching and False Discovery Rate Considerations

All discussed fragmentation methods produce consistent and recognizable fragment ions which may be used to determine a peptide sequence. The most common means of matching the fragment ions to a peptide sequence is by comparing the observed MS^2 spectrum to various *in silico* predicted spectra (Figure 1.1, B). Fortunately, many algorithms (e.g. Mascot[36], OMSSA[37] and SEQUEST[13]) can achieve this goal in an automated fashion. Theoretical spectra are generated from a protein database (e.g. *S. cerevisiae*) using various constraints, including minimum peptide length, peptide mass range, and enzymatic cleavage specify. As mentioned, accurate mass measurements reduce the number of these theoretical spectra which have to be considered when attempting to match them with an observed peptide spectrum. Spectral matching is a non-trivial endeavor, however, as observed MS^2 spectra vary

considerably in their quality and completeness of fragmentation. In addition observed MS^2 spectra may contain many unrelated ions (even fragment ions of unrelated co-eluting peptides), not considered in the theoretical spectra which negatively affect the matching process. Poor quality spectra tend to be miss-assigned, which increases the false discovery rate of the analysis. As such, these algorithms employ quality scoring metrics, such as the cross-correlation score in SEQUEST (XCorr) to compare various peptide spectral matches. These scores, along with peptide properties such as charge state, mass error, enzymatic cleavages specificity (if none was specified in the search), and others are useful for controlling the false discovery rate of a proteomic analysis.

The most widely used tool for estimating false discovery rates in a proteomics data set, and therefore the means to filter to a known false discovery rate, is achieved using the reverse database strategy[15] (Figure 1.2). Observed MS^2 spectra are searched against a composite database containing all proteins in their forward and reverse orientations (often referred to as the decoy database). Poor quality spectra which do not readily match a forward theoretical peptide sequence will randomly match forward or reverse sequences at an equal frequency[15]. Thus, the number of reverse hits is equal to half the total false positive peptide spectral matches (FDR = 2 * reverse hits/total hits). Other methods exist, including using a randomized sequence database[38], though the reverse database strategy in preferred for several reasons: The reverse database method preserves the structure of the forward database (same number of peptides considered in both forward and reverse orientation, same distribution of K/R and peptide lengths) and also preserves any sequence redundancy between proteins (such as between isoforms for example). The number of reverse hits, and their properties (XCorr, mas error, etc.) can be used as a guide for filtering a data set to a known false discovery rate.

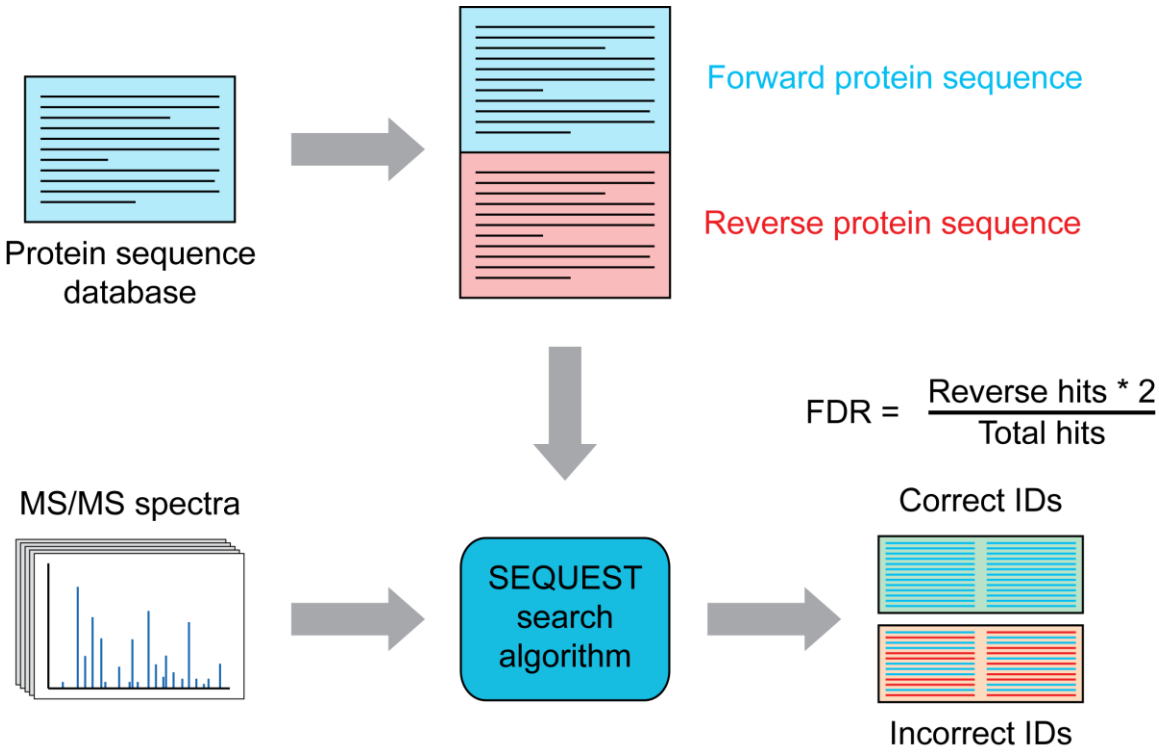


Figure 1.2. The reverse database strategy for false discovery rate estimation. The first step in this strategy is to generate a decoy database from which known false positive hits can be observed. Though several methods exist for obtaining a decoy database, the reverse database strategy is applied most widely. This method involves reversing the sequences of a protein database, so that the N- and C-termini of all tryptic peptides, from example, are inverted. The original forward and this generated reverse database are concatenated and used as the new database for a SEQUEST search. Peptide identification errors have an equal chance of matching a random forward or reverse peptide sequence. With this in mind, the false discovery rate (FDR) is estimated as twice the number of reverse hits, divided by the total number of hits. This strategy is also applicable for protein level false discovery rate estimations. The number of reverse hits is used as a guide for filtering incorrect assignments from a dataset while in parallel controlling the FDR of the dataset.

A means of separating false positive hits (forward and reverse) from true positive (forward only) hits, in order to control the false discovery rate, is through linear discriminant analysis[18] (LDA, Figure 1.3). LDA in general is used to identify features which can separate two of more data classes, which in the context of the reverse database strategy signifies the forward and reverse hits. Many of the discussed peptide properties are used as features in LDA (e.g. XCorr, mass error, etc.). Linear discriminant models are calculated for each run using peptide matches to forward and reversed protein sequences as positive and negative training data. Covariance between features and the mean values of these features from the positive and negative data sets are used to calculate the coefficients for

discriminant score calculations; a score is calculated for each peptide. After completion of LDA, a clear separation generally exists among true positive and false positive hits; indeed two distributions of peptides are seen, one containing true positive forward hits and one containing forward and reverse false positive hits (Figure 1.3, A). Peptides in each MS/MS run are ranked by descending discriminant score and filtered to a known false discovery rate (generally 1%) based on the number of reverse sequences remaining in the data set (Figure 1.3, B). By comparing the distributions of the true and false positive peptides (Figure 1.3, A), the probability that a given peptide has been incorrectly assigned can be estimated by its discriminant score. Other comparable methods employ the use of support vector machine (SVM) algorithms to define the separation between forward and reverse hits and yield comparable results[39].

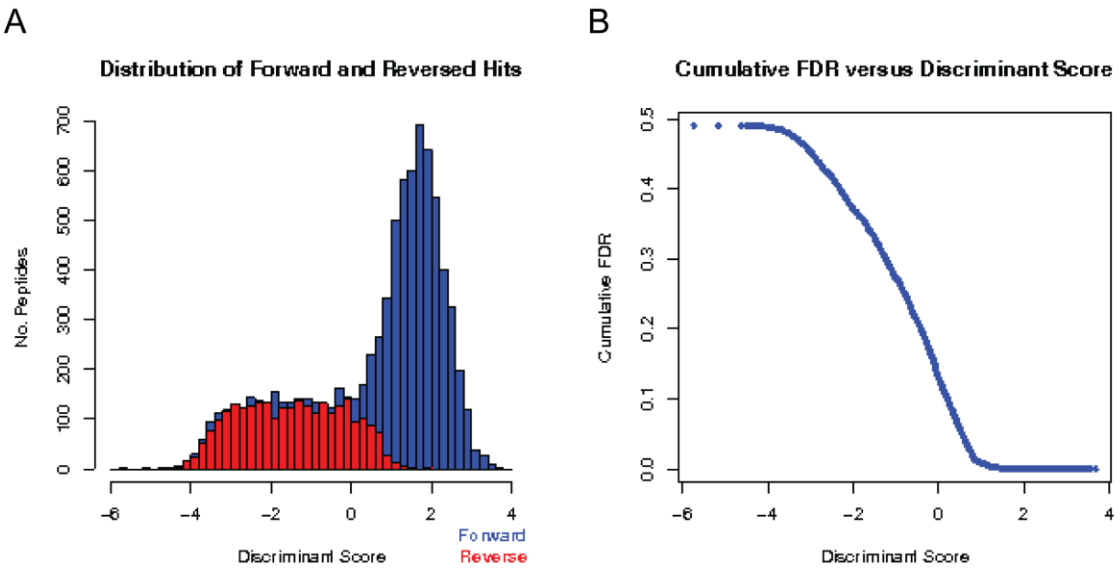


Figure 1.3. LDA analysis and peptide false discovery rates. (A) Linear discriminant scoring effectively separates false positive and true positive hits. Two distributions are observed: A large, sharp distribution of true positive hits (forward database matches) is centered at a discriminant score of ~2. A second diffuse distribution of false positive hits (forward and reverse database matches) is centered at a discriminant score of ~-2. Importantly false positive hits from the forward and reverse database display a similar distribution. (B) Discriminant scoring effectively controls the false discovery rate of an analysis. Peptides can be sorted by their discriminant score, and filtered to a known FDR (e.g. 1%), based on the number of reverse hits remaining at a given discriminant score.

Data sets are subsequently filtered to control the protein level false discovery rate. Scores are first created for each protein by multiplying the LDA peptide probabilities of all unique peptides from that protein. The resulting values are sorted by rank, and filtered to 1% FDR as described for peptides above, based on the presence of reverse hits[18]. Similar metrics have been applied in other proteomic studies[40-42]. Controlling the protein level false discovery rate is demonstrated to be important for controlling the total data set false discovery rate in chapter three; combining multiple LC-MS analyses, even with peptide level data filtering, drastically increases protein level false discovery rates, as often false positive hits are assigned to different proteins. Any time a group of samples is simultaneously considered, new filters must be created to properly assess the protein level false discovery rates. Such additional filtering is also required for limiting the false discovery rate of posttranslational modifications, and is an important consideration with MS^1 based quantification (detailed in chapter 3).

Analysis of Posttranslational Modifications

In addition to the discussed procedure which is generally applicable to peptide analysis by shotgun sequencing, posttranslational modifications, particularly phosphorylation, contain additional analytical considerations. The most obvious consideration is the abundance of phosphorylation. Although it is a ubiquitous modification, involved in virtually every biological process, phosphorylation events are generally of low stoichiometry[32]. In many cases the phosphorylated version of a protein may represent less than one percent of the total protein[32]. Since the depth of a shotgun sequencing analysis is affected by the dynamic range of a mixture, the identification of phosphopeptides requires additional enrichment steps prior to LC-MS/MS. Common enrichment techniques include strong cation exchange (SCX) chromatography, and immobilized metal affinity chromatography (IMAC)[43].

SCX separates peptides based on their solution charge state, and the addition of a phosphate group affects a peptide's solution charge. Under the acidic conditions of SCX, histidine, lysine and

arginine residues are protonated and therefore carry a positive charge, whereas phosphate moieties are deprotonated and carry a negative charge. Non-phosphorylated tryptic peptides tend to have +2 solution charges[44] (due to a protonated N-terminus and a protonated Lys/Arg). The addition a phosphate group will reduce the solution charge state by one to +1. This effect allows SCX to separate phosphorylated peptides from unmodified peptides. Often, however, many non-phosphorylated peptides are still contained within most SCX fractions, and further enrichment is required.

IMAC is a common method for such supplementary enrichment. IMAC resins used for the enrichment of phosphopeptides contain chelated iron, Fe (III). The negative phosphate group of a phosphopeptide is coordinated to the positively charged iron, thereby achieving enrichment. Enriched peptides are washed, and then released with phosphate salt treatment. A drawback of IMAC is that the resin also binds peptides carrying the negatively charged amino acids, aspartic and glutamic acid. A comparison of IMAC to titanium dioxide (TiO_2), a novel method for phosphopeptide enrichment at the time of its publication, is discussed in the next chapter. Since the publication of that chapter, TiO_2 using a lactic acid competitor prior to SCX has become the preferred phosphopeptide enrichment strategy[45]. Beyond these enrichment considerations, other technical hurdles exist in the analysis of phosphopeptides by LC-MS.

Due to the addition of a polar phosphate moiety, phosphopeptides tend to elute earlier during reverse phase separation, compared to their non-phosphorylated counterparts[46]; phosphopeptides also have been observed to elute over a smaller range of retention times[46]. Phosphorylation can affect the ability of trypsin and other proteases to cleave within the vicinity of that phosphorylated residue[47], the result of which is more missed cleavage events, creating longer peptides. As such lower initial concentrations of organic solvent are used to allow polar peptides to bind, and longer gradients are often applied to the analysis of phosphopeptide samples by LC-MS. Mass spectrometry components of the analysis are also affected by the phosphate group.

The fragmentation of phosphopeptides by CID, for example, is affected by the inclusion of a phosphorylated residue[48]. As discussed in the section on peptide fragmentation, the phosphate ester bond on modified serine and threonine residues is quite labile. When these peptide ions are subjected to fragmentation by CID, cleavage of this bond is the preferred fragmentation pathway. This preference reduces the intensity of sequence specific fragment ions generated through peptide backbone fragmentation. The result of this behavior is the generation of MS/MS spectra which are of poorer quality compared to their non-phosphorylated counterparts. This reduced quality affects the ability to correctly assign the MS^2 spectra. The acquisition of high mass accuracy precursor ion information, however, can compensate for the low quality MS^2. With such accuracy, spectral matching using automated database searching algorithms yields more correct assignments.

When multiple phosphorylatable residues exist on a peptide, correct site assignment is required. Unfortunately, neutral loss destroys much of the information pertaining to the location of a phosphorylation site. Generally, however, some site information is retained, albeit at lower intensities. As manual validation of site location is difficult in such cases and not feasible for large data sets, site scoring algorithms, such as the Ascore[49] are applied. Database searching methods such as SEQUEST do not contain such functionality. The Ascore algorithm confidently identifies site determining ions, which uniquely localize a site to a given residue, using a binomial probability model. In cases where insufficient site determining ions are present, the algorithm will localize a site to a peptide region. Such considerations are important, and affect both the false discovery rate of the analysis and site quantification. The effect of data set filtering on the false discovery rates of quantitative phosphoproteomic analyses is discussed in chapter three. An example of a typical phosphopeptide MS/MS spectrum is presented in Figure 1.4.

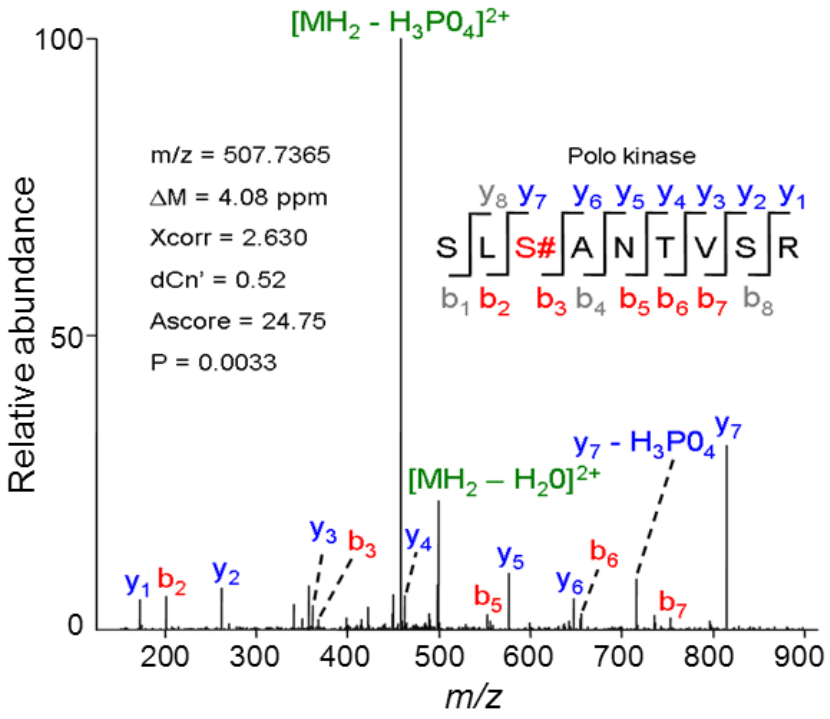


Figure 1.4. A typical MS/MS spectrum of a phosphopeptide with prominent neutral loss. An MS/MS spectrum of the peptide SLS#ANTVSR (2+), where 'S#' is the phosphorylated serine residue, is presented. Two of the most prominent peaks are the neutral loss of phosphate and water. Most of the observed fragment ions are present, albeit at low intensity. Despite this neutral loss, an Ascore of ~25 (P<0.005) was obtained, demonstrating the ability of the ion trap to detect relevant low abundance ions.

Quantitative Analysis by LC-MS and LC-MS/MS

Peptide quantification can be achieved either based on data from MS^1 or MS^2 spectra, depending on the quantitative method of choice. In both cases the ability to condense multiple experiments into a single LC-MS analysis, quantitative multiplexing, is presented. The common theme among all quantification methods is the introduction of stable isotopes (e.g. ^{13}C, ^{15}N, and ^{2}H) into peptides, so that they may be differentially detected during *m/z* analysis. A useful property of stable isotopes is that they are chemically identical, and hence consideration of chromatography (though deuterium isotopes can exhibit a chromatographic effect), ionization, fragmentation and other relevant concerns are equal between isotopomers. The advantages and limitations of the different strategies are discussed below.

Within MS^1 (precursor ion) based quantification, several methods are available (Figure 1.5). The most recognized means of peptide quantification is metabolic incorporation of stable isotopes (e.g. SILAC), which generally occurs by growing cells in media containing light (natural) or heavy (^{13}C and/or ^{15}N containing) amino acids. Cells which are grown under the heavy condition will incorporate the heavy amino acids into the newly synthesized proteins. When the light and heavy samples are mixed and analyzed by LC-MS, two isotopically related groups of peaks are observed. Although metabolic incorporation of stable isotope has been quite successful, and is required for certain experiments (e.g. analysis protein turnover via pulse-chase experiments) it is not generally feasible for quantitative proteomic analysis of larger animals. Problems with metabolic labeling in larger animals include the cost of labeling, experimental timing, isotopic enrichment, and others. An alternative to metabolic labeling is the incorporation of stable isotopes through chemical derivatization.

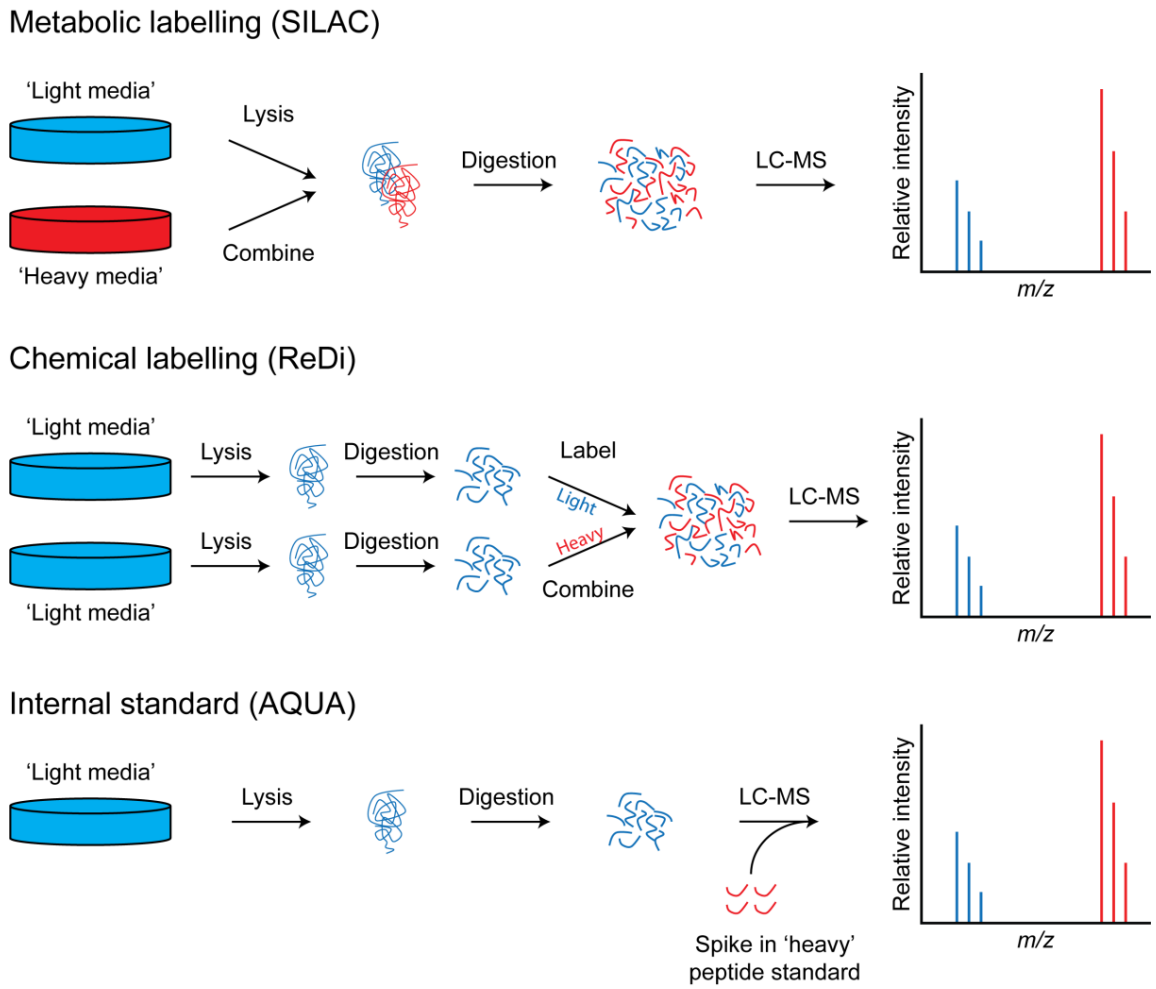


Figure 1.5. Different strategies for MS^1 based quantification. All methods employ incorporation of stable isotope into proteins or peptides to obtain quantitative information among the analytes. Three major strategies for MS^1 based quantification exist: metabolic incorporation of stable isotopes (e.g. SILAC), chemical incorporation of stable isotopes (e.g. reductive dimethylation, ReDi) and an internal standard approach (e.g. AQUA). With metabolic incorporation, stable isotopes are directly introduced into proteins by the inclusion of ^{13}C and ^{15}N containing lysine and arginine in the media (the 'heavy' media). Cells in 'light media' are grown in media containing isotopically natural versions of these amino acids. Cells are combined, lysed, digested and analyzed by LC-MS to obtain quantitative information. Chemical labeling approaches incorporate stable isotopes into peptides through chemical reactions which utilize isotopically related reagents. In the example of ReDi, heavy reaction occurs with ^2H containing formaldehyde (CD_2O) and sodium cyanoborodeuteride ($NaBD_3CN$). The light reaction occurs with the isotopically natural versions. In this method both conditions are grown in light media, each condition is separately lysed and, digested and labeled with the appropriate reagents. The labeled peptides are combined and analyzed by LC-MS to obtain quantitative information in the same manner as metabolic labeling. Quantification by internal standard involves spiking in a known amount of a heavy synthetic peptide (or a pool of heavy peptides) into a sample prior to LC-MS. The sequence of the heavy peptide is identical to a peptide of interest within the sample (of which phosphopeptides may be included). The ratio of this heavy peptide to the light peptide of interest can be used to calculate the absolute abundance of the light peptide. In this strategy, light samples are prepared in the standard manner, prior to the addition of a heavy internal standard. In all cases the mass differences between heavy (stable isotope containing) and light (natural isotope abundances) peptide can be observed in the MS^1 scan. Successive MS^1 scans are used to generate extracted ion chromatograms for quantification as discussed in Figure 1.6.

A common method for protein derivatization is via isotope coded affinity tags (ICAT)[50]. In this method, proteins are modified (e.g. on cysteine residues) by chemical tags which contain either natural isotopic abundances of hydrogen, or those which contain deuterium (d8). Heavy and light labeled samples are combined and digested. These tags also contain a biotin group, which can be used to purify modified peptides using avidin affinity chromatography. Purified peptides are analyzed by LC-MS/MS. An alternative method for the chemical incorporation of stable isotopes into peptides is reductive dimethylation (ReDi). This method is the subject of chapter three. Peptides are reductively dimethylated with natural or $^{2}H/^{13}C$ containing formaldehyde and sodium cyanoborohydride or cyanoborodeuteride, and are quantified in the same manner as SILAC peptide pairs. One final means of MS^1 based quantification is the use of heavy internal standards for the absolute quantification of peptides (AQUA[8]). Prior to LC-MS analysis, a synthetic heavy version of a peptide of interest is spiked into the sample at a known concentration. The light version of the peptide is identified and its abundance is compared to this internal standard, to obtain its absolute quantification. One interesting application of this method is the assessment of kinase activity by LC-MS in a high throughput manner[51].

In all cases, the quantitative ratios between light and heavy species are obtained by comparing extracted ion chromatograms (XICs) for each differentially labeled form of a peptide (Figure 1.6). An XIC is created by obtaining peptide intensity information from the MS^1 scan across the elution profile of the peptide. The XIC of the heavy and light versions are integrated (area under the curve) and the comparison of these results yields the quantitative peptide ratio. Using this method, along with high mass accuracy, only one MS^2 scan (either heavy or light peptides version) is often required to obtain both sequence and quantitative information. High mass accuracy generally facilitates the correct identification of heavy and light species over time: with accurate mass, peptides which match the correct charge state, are within a small mass tolerance (<5 ppm), and within a short time window of the identified peptide are likely the ions of interest. Such quantification is greatly aided by accurate

precursor mass measurements. Peptides are collapsed into proteins, and generally the median quantitative peptide ratio is used as the reported protein ratio.

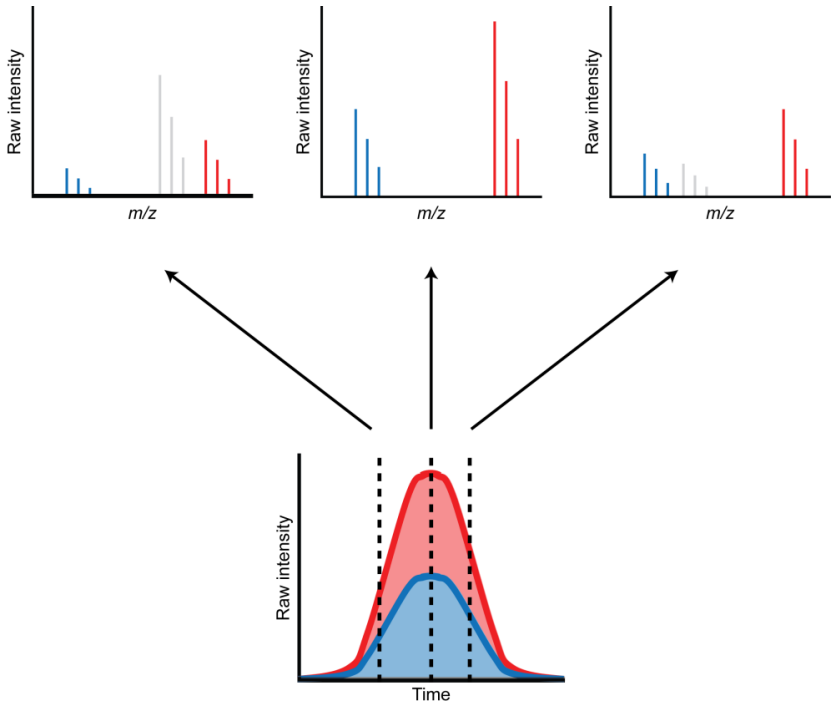
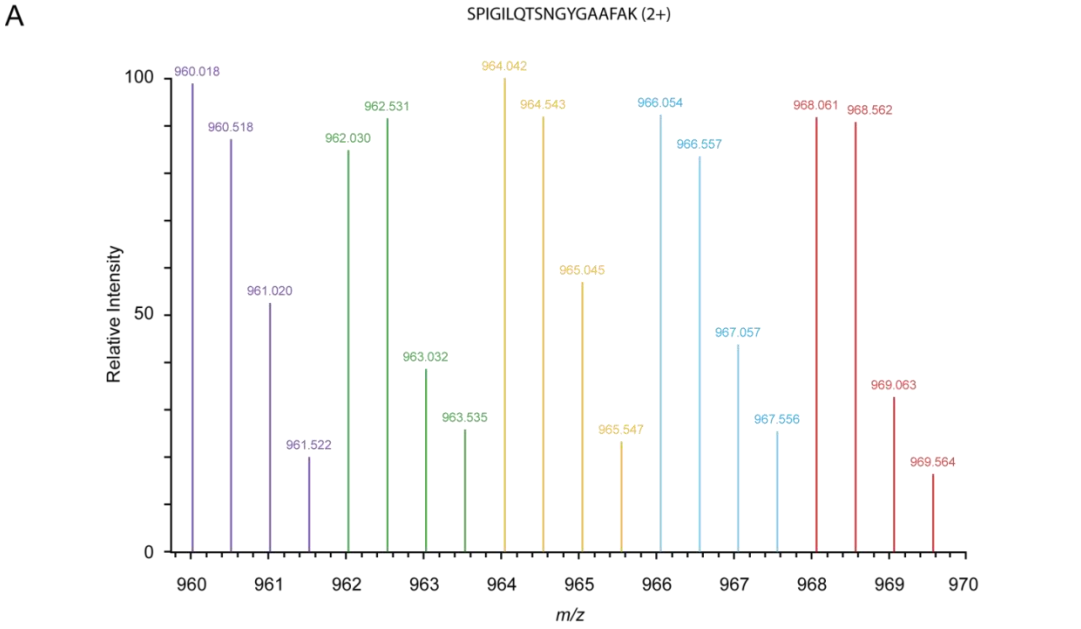


Figure 1.6. Peptide quantification using extracted ion chromatograms (XICs). As each MS^1 scan only provides a snapshot of a peptide's elution, and thus only represents a fraction of its total abundance, extracted ion chromatograms are preferred for accurate MS^1 based quantification. Once a peptide's identification is obtain from an MS^2 scan, quantitative information is extracted over the elution profile for the heavy and light versions of a peptide, which defines the XIC. In example above, peptide peaks from MS^1 survey scans at ~1/3 maximal, maximal, and ~1/2 maximal intensity of the extracted ion chromatogram are displayed. Light peptide peaks displayed in blue, heavy peptides peaks are displayed in red, and unrelated peptide peaks are displayed in gray. An important consideration is that mass accuracy allows for the resolution of the desired heavy and light peaks, so that unrelated peak information does not affect quantification. Integration of the light and heavy XICs provides area under the curve, from which the heavy to light peptide ratios are obtained.

Interestingly, one can use many different forms of stable isotopes simultaneously, creating multiple isotopically related peptides for quantitative analysis. Quantitative multiplexing using precursor ion (MS^1) quantification is particularly accessible by means of reductive dimethylation, as differential mass additions of 2 Da are easily obtained through various reagent combinations (Figure 1.7). Using reductive dimethylation, five different samples may be simultaneously compared in one LC-MS analysis. One caveat of MS^1 based quantification, which is particularly true for MS^1 based quantitative multiplexing, is that additional peptide isotopic forms leads to a higher proteome complexity. This

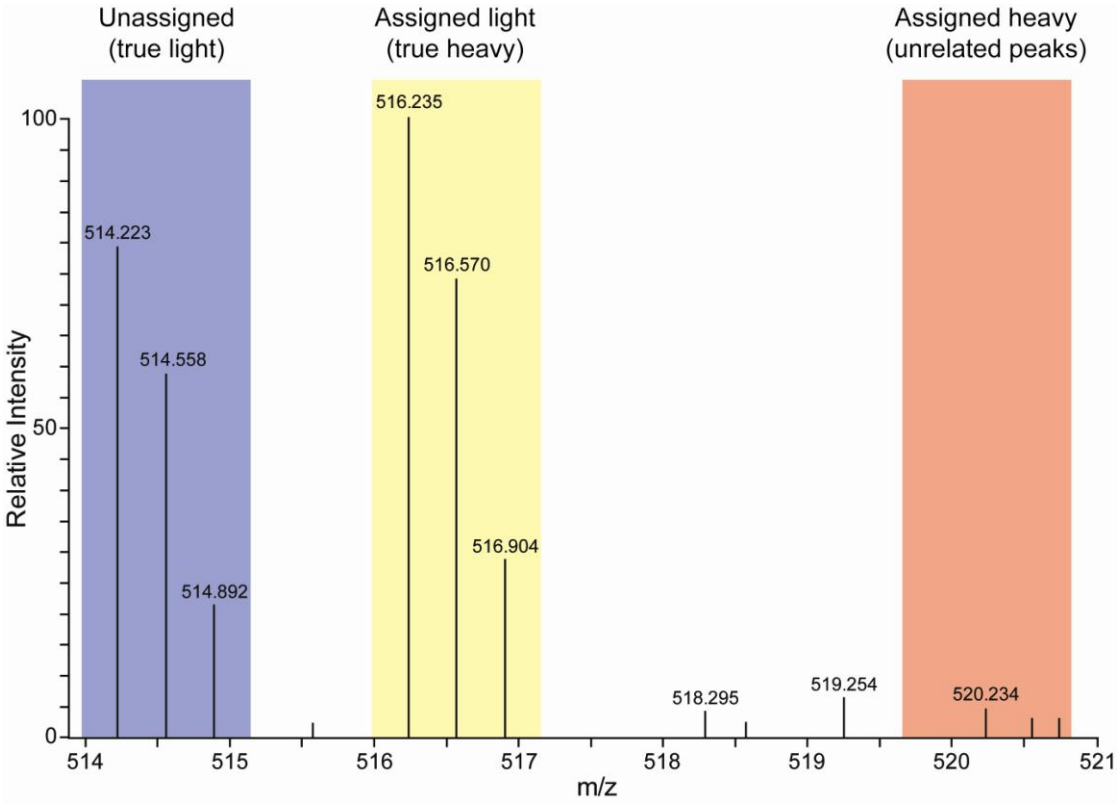
complexity, due to the stochastic nature of shotgun proteomics, leads to fewer peptide identifications. It is difficult to determine which peptide form has been selected for MS/MS while data is still in acquisition. Therefore, it is not trivial to strategically trigger MS/MS on only one isotopic form, in order to increase analytical depth. Such dynamic exclusion methods are an active area of research. This type of multiplexing may still be useful for less complex proteome analyses, such as immunoprecipitations.



Label	Formaldehyde	Cyanoborohydride	Mass shift (Da)
Unlabeled	N/A	N/A	0
$-(CH_3)_2$	CH_2O	$NaBH_3(CN)$	+28.0313
$-(CH_2D)_2$	CH_2O	$NaBD_3(CN)$	+2.0125
$-(CHD_2)_2$	CD_2O	$NaBH_3(CN)$	+4.0251
$-(CD_3)_2$	CD_2O	$NaBD_3(CN)$	+6.0376
$-(^{13}CD_3)_2$	$^{13}CD_2O$	$NaBD_3(CN)$	+8.0443

Figure 1.7. Reductive dimethylation supports up to five-plex quantification. (A) The MS^1 spectra of a five-plex reductively dimethylated yeast peptide is shown (SPIGILQTSNGYGAAFAK, 2+), displaying all five isotopically related forms. The displayed peptide contains two dimethyl labels (N-terminus and lysine); each successive isotopomer displays a 2 m/z shift with respect to the previous one (4 Da/2 charges). In this example, the peptides were mixed at a 1:1 ratio for all reaction conditions. (B) A table containing the combinations of formaldehyde, sodium cyanoborohydride and their heavy isotopes, required to create five-plex labeling are given. The label type and mass shift between isotopomers (relative to light) is also displayed in the aforementioned table. (C) The extracted ion chromatograms for this peptide are displayed; despite the presence of five isotopic forms, the elution profile for each species is readily obtained, with all isotopic species displaying similar elution profiles. All forms of the peptide display close to the expected 1:1 ratio with respect to one another.

Beyond consideration of proteomic depth, MS^1 based quantitative analysis has a much larger problem, in that a peptide's quantification is tied to its identification. The result of this behavior is that false peptide identifications often leads to erroneous quantification (Figure 1.8). A false positive hit in the forward database may be erroneously assigned as a light peptide (based on the matched sequence for a given parent ion *m/z*), when in reality it is a heavy peptide. In this example, no peak or an unrelated peak will be selected as its isotopomer when creating XICs. The ratio of the heavy and light XIC will often be much greater than 1:1 in these cases. Indeed the data which is considered regulated (e.g. 2-fold change) in MS^1 based quantification studies are enriched for false positive identifications, demonstrating the need for additional filtering. Ambiguity in site localization during quantitative phosphoproteomics further compounds these errors. The identification of this problem and functional solutions are discussed in chapter three. An exciting alternative to MS^1 based quantification which avoids these discussed problems is the use of isobaric (same nominal mass) reagents for MS^2 based quantification.



False quantification, ~ 0.01:1

m/z	Isotopomer	Charge state	Sequence	Number of labels
516.235	Light	3	ASTPS*QVNGIT*GAK	2 (N-term and K)
520.260	Heavy	3	A]STPS*QVNGIT*GAK#	2 (N-term and K)

True quantification, ~ 1:1

m/z	Isotopomer	Charge state	Sequence	Number of labels
514.223	Light	3	XXXXXXXXXXXXXXXR	1 (N-term only)
516.235	Heavy	3	X]XXXXXXXXXXXXXXR	1 (N-term only)

Figure 1.8. **MS^1 based quantification errors often occur for incorrectly assigned peptide ions due the misassignment of heavy and light peptide ions.** The presented peptide spectral match (observed m/z of 516.235, 3+) was assigned to the sequence ASTPS*QVNGIT*GAK, where * represents a phosphorylated residue. This assignment, however, appears to be erroneous, due to the presence of a light isotopomer of the 516 Th peak (observed m/z 514.223, +3), and the absence of a likely heavy isotopomer of the assigned peptide (theoretical m/z 520.260, 3+). The result of this misassignment is a large (>100 fold) calculated ratio between the light and heavy species, whereas the likely true quantitative ratio for the correctly assigned peptide is closer to 1:1. The likely correct peptide assignment contains an arginine tryptic cleavage site at the c-terminus of the peptide, and thus contains the dimethylation label on the n-terminus only. Isotopomer misassignment and the resulting erroneous quantification contribute to the observation that false positive peptides are preferentially distributed amongst the "regulated peptide" data set (e.g. those changing by two-fold).

Common reagents which are used in MS^2 based quantification are tandem mass tags (TMT) [52, 53] and isobaric tags for relative and absolute quantitation (iTRAQ) [54]. In this dissertation, only TMT was used. TMT exists as six related isobaric reagents which tag free amine group. Each reagent is used to label peptides from a biological sample of interest (treatment vs. no treatment, mutants etc.). These reagents contain a linker region and a reporter ion region with a cleavage site in between. The isobaric nature of these reagents is a result of balancing stable isotopes (^{13}C and ^{15}N) among the linker and reporter ion regions so that the intact reagents are all the same mass. As such, the inclusion of six isotopically related forms does not increase signal complexity in the MS^1 spectra, avoiding one drawback of MS^1 based multiplexing. Upon HCD fragmentation the cleavage site is broken, and the differential mass characteristics of these reagents is observed in the MS^2 spectrum (Figure 1.9). The intensity distribution of low mass reporter ions is used for the quantification of peptides between biological conditions. As with MS^1 based quantification, the TMT reagents are chemically identical, and thus behave identically in each step of the sample preparation procedure such as chromatographic separation.

Quantification occurs in the MS^2 spectrum, and is independent of a peptide's identification. In a given experiment, the majority of peptide analytes are at 1:1 ratios with respect to one another. Hence, even if a peptide is assigned a fallacious sequence, the quantitative ratio observed among TMT channels will be accurate and close to 1:1. In this manner, TMT avoids the false discovery issues of MS^1 based quantification. Indeed no correlation is observed between known false positive identifications (reverse hits) and regulated protein data (e.g. 2-fold change). Quantitative multiplexing is the subject of chapter four, where the drawbacks and solutions regarding TMT quantification are discussed in detail. In this chapter common proteome-wide experimental types are presented, and the means of analyzing the complex data generated from these experiments is highlighted.

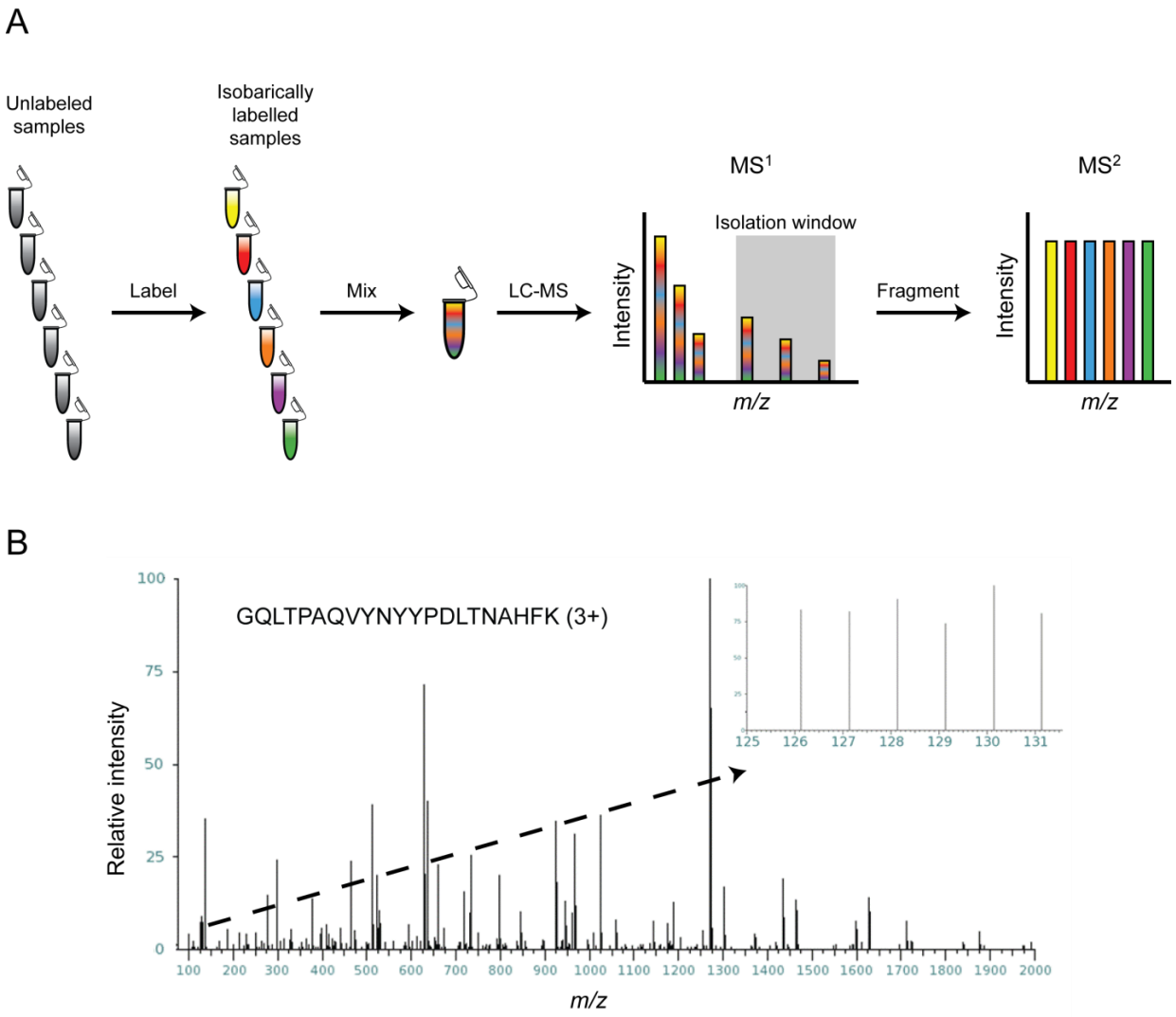


Figure 1.9. Typical strategy for MS^2 based multiplex quantification. (A) Up to six sample may be simultaneously analyzed using TMT. Due to the isobaric nature of the reagents, multiple samples can be labeled and combined without increased signal complexity in the MS^1 spectra. As with other stable isotope methods relying on ^{13}C and ^{15}N incorporation, the TMT reagents are chemically identical and indistinguishable based on chromatographic separation. Additionally, in a full MS spectrum, a peptide labeled with any of the six TMT reagents will have the same mass to charge ratio (m/z), thus maintain the complexity of an unlabeled samples. This behavior contrasts the increase in sample complexity observed with MS^1 quantification methods, such as SILAC and ReDi. Only once TMT labeled peptides are fragment by HCD, are the reporter ions generated. Peptide ratios (and thus protein ratios) are determined by the S/N ratios of the reporter ions. (B) Low mass reporter ions are visible along with the typical b- and y- type fragment ions in an MS^2 spectrum.

References

1. Sanger, F.; Coulson, A. R., A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **1975,** 94, (3), 441-8.
2. Sanger, F.; Air, G. M.; Barrell, B. G.; Brown, N. L.; Coulson, A. R.; Fiddes, C. A.; Hutchison, C. A.; Slocombe, P. M.; Smith, M., Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **1977,** 265, (5596), 687-95.
3. Venter, J. C., *et al.*, The sequence of the human genome. *Science* **2001,** 291, (5507), 1304-51.
4. Lander, E. S., *et al.*, Initial sequencing and analysis of the human genome. *Nature* **2001,** 409, (6822), 860-921.
5. Gasch, A. P.; Spellman, P. T.; Kao, C. M.; Carmel-Harel, O.; Eisen, M. B.; Storz, G.; Botstein, D.; Brown, P. O., Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **2000,** 11, (12), 4241-57.
6. Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R., Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **1999,** 19, (3), 1720-30.
7. Church, G. M.; Kieffer-Higgins, S., Multiplex DNA sequencing. *Science* **1988,** 240, (4849), 185-8.
8. Gerber, S. A.; Rush, J.; Stemman, O.; Kirschner, M. W.; Gygi, S. P., Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* **2003,** 100, (12), 6940-5.
9. Edman, P., A method for the determination of amino acid sequence in peptides. *Arch Biochem* **1949,** 22, (3), 475.
10. Peng, J.; Gygi, S. P., Proteomics: the move to mixtures. *J Mass Spectrom* **2001,** 36, (10), 1083-91.
11. Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **2001,** 19, (3), 242-7.
12. Haas, W.; Faherty, B. K.; Gerber, S. A.; Elias, J. E.; Beausoleil, S. A.; Bakalarski, C. E.; Li, X.; Villen, J.; Gygi, S. P., Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Mol Cell Proteomics* **2006,** 5, (7), 1326-37.
13. Eng, J. K.; McCormack, A. L.; Yates Iii, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **1994,** 5, (11), 976-989.
14. Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S., Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res* **2008,** 7, (1), 29-34.
15. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol Biol* **2007,** 604, 55-71.
16. Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P., Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* **2003,** 2, (1), 43-50.
17. Kim, W.; Bennett, E. J.; Huttlin, E. L.; Guo, A.; Li, J.; Possemato, A.; Sowa, M. E.; Rad, R.; Rush, J.; Comb, M. J.; Harper, J. W.; Gygi, S. P., Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol Cell* **2011,** 44, (2), 325-40.
18. Huttlin, E. L.; Jedrychowski, M. P.; Elias, J. E.; Goswami, T.; Rad, R.; Beausoleil, S. A.; Villen, J.; Haas, W.; Sowa, M. E.; Gygi, S. P., A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **2010,** 143, (7), 1174-89.
19. Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P., Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Methods* **2005,** 2, (9), 667-75.

20. Siuti, N.; Kelleher, N. L., Decoding protein modifications using top-down mass spectrometry. *Nat Methods* **2007,** 4, (10), 817-21.
21. Regnier, F. E.; Riggs, L.; Zhang, R.; Xiong, L.; Liu, P.; Chakraborty, A.; Seeley, E.; Sioma, C.; Thompson, R. A., Comparative proteomics based on stable isotope labeling and affinity selection. *J Mass Spectrom* **2002,** 37, (2), 133-45.
22. Karas, M.; Bachmann, D.; Hillenkamp, F., Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Analytical Chemistry* **1985,** 57, (14), 2935-2939.
23. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989,** 246, (4926), 64-71.
24. Wilm, M.; Mann, M., Analytical properties of the nanoelectrospray ion source. *Anal Chem* **1996,** 68, (1), 1-8.
25. Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M., Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **1996,** 379, (6564), 466-9.
26. Syka, J. E.; Marto, J. A.; Bai, D. L.; Horning, S.; Senko, M. W.; Schwartz, J. C.; Ueberheide, B.; Garcia, B.; Busby, S.; Muratore, T.; Shabanowitz, J.; Hunt, D. F., Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J Proteome Res* **2004,** 3, (3), 621-6.
27. Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S., Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom Rev* **1998,** 17, (1), 1-35.
28. Makarov, A., Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem* **2000,** 72, (6), 1156-62.
29. Bakalarski, C. E.; Elias, J. E.; Villen, J.; Haas, W.; Gerber, S. A.; Everley, P. A.; Gygi, S. P., The impact of peptide abundance and dynamic range on stable-isotope-based quantitative proteomic analyses. *J Proteome Res* **2008,** 7, (11), 4756-65.
30. Michalski, A.; Damoc, E.; Lange, O.; Denisov, E.; Nolting, D.; Muller, M.; Viner, R.; Schwartz, J.; Remes, P.; Belford, M.; Dunyach, J. J.; Cox, J.; Horning, S.; Mann, M.; Makarov, A., Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol Cell Proteomics* **2011,** 11, (3), O111 013698.
31. Michalski, A.; Damoc, E.; Hauschild, J. P.; Lange, O.; Wieghaus, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S., Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol Cell Proteomics* **2011,** 10, (9), M111 011015.
32. Wu, R.; Haas, W.; Dephoure, N.; Huttlin, E. L.; Zhai, B.; Sowa, M. E.; Gygi, S. P., A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nat Methods* **2011,** 8, (8), 677-83.
33. Mikesh, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E.; Shabanowitz, J.; Hunt, D. F., The utility of ETD mass spectrometry in proteomic analysis. *Biochim Biophys Acta* **2006,** 1764, (12), 1811-22.
34. Lee, M. V.; Topper, S. E.; Hubler, S. L.; Hose, J.; Wenger, C. D.; Coon, J. J.; Gasch, A. P., A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Mol Syst Biol* **2011,** 7, 514.
35. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M., Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* **2007,** 4, (9), 709-12.
36. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999,** 20, (18), 3551-67.
37. Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H., Open mass spectrometry search algorithm. *J Proteome Res* **2004,** 3, (5), 958-64.

38. Higdon, R.; Hogan, J. M.; Van Belle, G.; Kolker, E., Randomized sequence databases for tandem mass spectrometry peptide and protein identification. *OMICS* **2005,** 9, (4), 364-79.
39. Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J., Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **2007,** 4, (11), 923-5.
40. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **2003,** 75, (17), 4646-58.
41. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **2008,** 26, (12), 1367-72.
42. Wenger, C. D.; Phanstiel, D. H.; Lee, M. V.; Bailey, D. J.; Coon, J. J., COMPASS: a suite of pre- and post-search proteomics software tools for OMSSA. *Proteomics* **2011,** 11, (6), 1064-74.
43. Villen, J.; Gygi, S. P., The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat Protoc* **2008,** 3, (10), 1630-8.
44. Beausoleil, S. A.; Jedrychowski, M.; Schwartz, D.; Elias, J. E.; Villen, J.; Li, J.; Cohn, M. A.; Cantley, L. C.; Gygi, S. P., Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A* **2004,** 101, (33), 12130-5.
45. Kettenbach, A. N.; Gerber, S. A., Rapid and reproducible single-stage phosphopeptide enrichment of complex peptide mixtures: application to general and phosphotyrosine-specific phosphoproteomics experiments. *Anal Chem* **2011,** 83, (20), 7635-44.
46. Kim, J.; Petritis, K.; Shen, Y.; Camp, D. G., 2nd; Moore, R. J.; Smith, R. D., Phosphopeptide elution times in reversed-phase liquid chromatography. *J Chromatogr A* **2007,** 1172, (1), 9-18.
47. Benore-Parsons, M.; Seidah, N. G.; Wennogle, L. P., Substrate phosphorylation can inhibit proteolysis by trypsin-like enzymes. *Arch Biochem Biophys* **1989,** 272, (2), 274-80.
48. DeGnore, J. P.; Qin, J., Fragmentation of phosphopeptides in an ion trap mass spectrometer. *J Am Soc Mass Spectrom* **1998,** 9, (11), 1175-88.
49. Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P., A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* **2006,** 24, (10), 1285-92.
50. Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **1999,** 17, (10), 994-9.
51. Kubota, K.; Anjum, R.; Yu, Y.; Kunz, R. C.; Andersen, J. N.; Kraus, M.; Keilhack, H.; Nagashima, K.; Krauss, S.; Paweletz, C.; Hendrickson, R. C.; Feldman, A. S.; Wu, C. L.; Rush, J.; Villen, J.; Gygi, S. P., Sensitive multiplexed analysis of kinase activities and activity-based kinase identification. *Nat Biotechnol* **2009,** 27, (10), 933-40.
52. Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C., Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **2003,** 75, (8), 1895-904.
53. Dayon, L.; Hainard, A.; Licker, V.; Turck, N.; Kuhn, K.; Hochstrasser, D. F.; Burkhard, P. R.; Sanchez, J. C., Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags. *Anal Chem* **2008,** 80, (8), 2921-31.
54. Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **2004,** 3, (12), 1154-69.

Chapter 2

Phosphoproteome Analysis of Fission Yeast

Attributions:

- This chapter contains work published as Wilson-Grady, J. T., Villén, J., and Gygi, S. P., Phosphoproteome Analysis of Fission Yeast. J Proteome Res 2008, 7, (3), 1088-97.

- J.T. Wilson-Grady performed the experimental steps and data analysis, as well as manuscript preparation. All bioinformatic processing was performed using in-house software developed by the “GFY Development Team.”

- J. Villén guided the experimental steps and data analysis, and also edited the manuscript.

- S.P. Gygi advised the project.

Abstract

Phosphorylation is a key regulator of many events in eukaryotic cells. The acquisition of large-scale phosphorylation data sets from model organisms can pinpoint conserved regulatory inputs and reveal kinase-substrate relationships. Here we provide the first large-scale phosphorylation analysis of the fission yeast, *Schizosaccharomyces pombe.* Protein from thiabendazole-treated cells was separated by preparative SDS-PAGE and digested with trypsin. The resulting peptides were subjected to either IMAC or TiO_2 phosphopeptide enrichment methods and then analyzed by LC-MS/MS using an LTQ-Orbitrap mass spectrometer. In total, 2887 distinct phosphorylation sites were identified from 1194 proteins with an estimated false discovery rate of <0.5% at the peptide level. A comparison of the two different enrichment methods is presented, supporting the finding that they are complementary. Finally, phosphorylation sites were examined for phosphorylation-specific motifs and evolutionary conservation. These analyses revealed both motifs and specific phosphorylation events identified in *S. pombe* were conserved and predict novel phosphorylation in mammals.

Introduction

The fission yeast, *Schizosaccharomyces pombe*, has traditionally been used as a tool to study cell cycle regulation, particularly with respect to mitosis [1-3]. Along with the budding yeast *S. cerevisiae*, *S. pombe* is a proven model organism for studying mitosis-specific molecular events, notably because of its small genome and that it undergoes cell division in a similar fashion to higher eukaryotes. With the publication of the complete genome sequence by the Sanger Institute[4] and the global characterization of protein localization and expression levels [5], in depth biological analyses using fission yeast continue to expand its utility as a model organism. *S. pombe* shares as much homology with humans as it does with *S. cerevisiae*, with many proteins showing more similarity to mammalian homologs than to their homologs in *S. cerevisiae* [6, 7]; therefore, research on fission yeast may have direct impact on clinically relevant research in humans. Indeed, fission yeast contain numerous proteins with human homologs linked to diseases (>172 to date) such as diabetes, cancer, cystic fibrosis and recently Parkinson's disease[4, 8]. According to our current understanding, protein phosphorylation is the most important regulator of the cell cycle and many other processes commonly studied in *S. pombe* [9-12]. Despite this understanding, a relatively small number of phosphorylation events have been directly characterized in fission yeast.

Large-scale phosphorylation analysis by mass spectrometry is emerging as a powerful technique in signaling research. A common feature of all large-scale studies is the requirement for phosphopeptide enrichment due to the general low abundance of phosphorylated species. In fact, most studies to date providing >500 phosphopeptides have two features in common: protein (or peptide) fractionation and phosphopeptide enrichment [13-18]. Though the exception to this trend has been powerful antibody-based isolations which do not require pre-fractionation [13, 19, 20]. Commonly used fractionation techniques include SDS-PAGE separation and strong cation exchange (SCX)

chromatography. The two most common phosphopeptide enrichment methods use immobilized metal affinity chromatography (IMAC) and titanium dioxide chromatography (TiO_2). IMAC is based on affinity capture of phosphopeptides, whereas the TiO_2 method uses acid/base chemistry to selectively enrich for phosphorylated peptides [21, 22]. Recently, it has been reported that these methods are complementary and combining them provides an aggregate data set, larger than either single method by itself [23].

A long term goal of many groups is to provide large phosphorylation databases for many model organisms, tissues and cell lines as resources for understanding important molecular regulation events across several species. With the completion of several data sets from many species [13-18], we can begin to understand phosphorylation events in an evolutionary context, perhaps identifying key regulatory steps in varied biological processes. In this study, SDS-PAGE combined with IMAC/TiO_2 and LC-MS yielded 2887 distinct phosphorylation sites from 1194 proteins, the largest yeast phosphorylation data set to date. Additionally, we identified conserved phosphorylation events in fission yeast that persist in humans after 500 million years of evolution, validating the potential for these large data sets to predict novel phosphorylation events in other species, and to contribute to a deeper understanding of evolutionary history.

Materials and Methods

All chemicals not specified were obtained from Sigma-Aldrich, St. Louis, MO.

Thiabendazole treatment and protein extraction from *S. pombe*.

Fission yeast (wt strain 972h-), grown to an OD_{600} of 0.8, was treated for 3h with the microtubule depolymerizing agent thiabendazole at 25µg/mL (final concentration), a dose sufficient to ensure a high degree of metaphase arrest. Approximately 5 X 10^9 cells were pelleted (3000 rpm, 5 min,

4 ºC) and rinsed with 1 mL ice cold Milli-Q water, pelleted and flash frozen with liquid nitrogen. Total protein extracts were obtained by bead beating (Mini-BeadBeater 8, Biospec) the thawed cells in 400 µL of urea lysis buffer [50 mM Tris/75 mM NaCl/8 M urea (pH 8.0)/10 mM sodium pyrophosphate/1 mM sodium fluoride/1 mM β-glycerophosphate/1 mM sodium orthovanadate/1 tablet complete Mini protease inhibitor mixture (Roche) per 10 ml]. Bead beating was performed using excess glass beads (0.5 mm, Biospec) for three pulses of 45 s, at 4 ºC. The lysate was separated from the beads, and insoluble components were removed by centrifugation (14,000 rpm, 5 min, 4ºC). Protein concentration was determined using the Bradford method.

Disulfide reduction, preparative SDS-PAGE and in-gel proteolysis

Disulfide bonds were reduced with DTT (5mM, 56ºC, 45 min) and free sulfhydryl groups were alkylated with iodoacetamide in the dark (15mM, 25 ºC, 45 min). A hand-poured 10% acrylamide SDS-PAGE gel (15 X 15 X 0.15 cm, solution from Bio-Rad, Hercules, CA) was used to separate 6 mg of whole cell lysate. Six mg is at the upper limit of separation for preparative SDS-PAGE and was the amount used previously in our large-scale phosphorylation analysis of budding yeast[14]. Electrophoresis was stopped when the dye front reached 8 cm. The Coomassie-stained gel was excised into 12 regions (Figure 2.1), which were each further cut into ~1 mm cubes and transferred into 15 mL conical tubes. In-gel digestion was carried out as previously described [14], with the exception of peptide extractions, which occurred twice with 50% $(CH_3)_2CHOH$/5% CH_3COOH and then once with 50% CH_3CN/H_2O/5% HCOOH. Extracted peptides were split precisely into two aliquots (for subsequent steps), dried to completion by vacuum centrifugation and stored at -20ºC.

Phosphopeptide enrichment using IMAC

IMAC resin (Phos-Select iron affinity gel; Sigma, St. Louis, MO) was equilibrated with three washes of 250mM CH_3COOH/30% CH_3CN. Peptide samples were resuspended in 100 µL of 250mM CH_3COOH/30% CH_3CN and transferred to PCR tubes containing 10 µL of equilibrated IMAC slurry (1:1, beads:liquid, 5 µL beads). After a 60 min incubation (25 ºC with vigorous shaking), the supernatant was collected, and the resin was washed three times with 200 µL 250mM CH_3COOH/30% CH_3CN, adding each wash to the supernatant (non-phosphorylated peptides). The phosphopeptides were eluted from the resin with three 70 µl washes of 50 mM K_2HPO_4/NH_3, pH 10.0, into tubes containing 20 µl 10% formic acid. Eluted peptides, and flow thru collected from each gel band were dried by vacuum centrifugation, and subsequently desalted using C18 Empore Disks (3M Corporation, Minneapolis, MN [24]).

Phosphopeptide enrichment using TiO_2

The procedure for TiO_2 enrichment was adapted from Larsen et al [25]. TiO_2 (Titansphere, GL Sciences, Japan) slurry was prepared at a concentration of 25 mg/mL in 50% CH_3CN/0.1% trifluoroacetic Acid (TFA). 20 µL of slurry was added to a PCR tube for each sample and washed twice with 150 µL of 1.5M dihydroxybenzoic acid (DHB)/50% CH_3CN/0.1% TFA. The peptide samples were resuspended in 50 µL of 1.5M DHB/50% CH_3CN/0.1% TFA and added to the tubes of resin. After a 60 min incubation (25 ºC with vigorous shaking), the supernatant was discarded, and the resin washed once with 150 µL of 0.25M DHB/50% CH_3CN/0.1% TFA and twice with 150 µL of 50% CH_3CN/0.1% TFA. The phosphopeptides were eluted from the washed resin three times with 20 µL of 50 mM K_2HPO_4/NH_3, pH 10.8, into tubes containing 20 µl 50% CH_3CN/5 % HCOOH. As with the IMAC procedure, the samples were lyophilized by vacuum centrifugation and desalted using C18 Empore Disks.

LC-MS/MS analysis

All LC-MS/MS data were obtained using an LTQ-Orbitrap hybrid mass spectrometer (Thermo Fischer, San Jose, CA). Dried phosphopeptide enriched samples were resuspended in 6 µL of 5% CH_3CN/4% HCOOH, and 4 µL were loaded onto a pulled fused silica microcapillary column (125 µm, 18 cm bed volume) packed with C_{18} reverse-phase resin (Magic C18AQ; 5-µm particles; 200-Å pore size; Michrom Bioresources, Auburn, CA) using a Famos autosampler (LC Packings, San Francisco, CA). Once loaded, the phosphopeptides were separated using an Agilent 1100 series binary pump across a 40 min linear gradient of 6% to 24% CH_3CN in 0.125% HCOOH at a flow rate of 600 nl/min. The IMAC flow-through samples were separated in a similar manner using a 66 min linear gradient, due to the increased complexity of the samples. In each data collection cycle, one full MS scan (375-1800 *m/z*) was acquired in the Orbitrap ($6x10^4$ resolution setting, automatic gain control (AGC) target of 10^6), followed by 10 data-dependent MS/MS scans in the LTQ (AGC target, 5,000; threshold 3,000) using the 10 most abundant ions and collision-induced dissociation (CID) for fragmentation[26]. The method dynamically excluded previously selected ions for 30 s, as well as rejected singly charged ions and unassigned charge states.

Database searching

RAW files obtained from data collection were converted into mzXML format using the ReAdW program (http://sashimi.sourceforge.net/software_glossolalia.html). Monoisotopic precursor ion and charge state information for each acquired MS/MS spectrum were corrected by in-house software. The SEQUEST search algorithm (version 27, revision 12) was used to search MS/MS spectra against a composite database comprised of all the *S. pombe* ORFs (downloaded from the Sanger Institute[4] on 12/08/2006, ftp://ftp.sanger.ac.uk/pub/yeast/pombe/Protein_data/pompep) in their forward and reversed orientations [27]. The search parameters used for posttranslational modification included a static

modification of 57.02146 Da on cysteine (carboxyamidomethylation), and dynamic modifications of 15.99491 Da on methionine (oxidation) and 79.96633 Da on serine, threonine, and tyrosine (phosphorylation).

Data filtering and phosphorylation site localization

Data were filtered to contain less than 1% false positives estimated by the number of decoy matches[27], using in house software and filtering based on dCn' (previously defined[13]), XCorr, charge state, mass error, and phosphorylation. Optimized thresholds were further refined for XCorr, peptide score and mass error in a similar manner as previously described[14]. After removing decoy database-derived phosphopeptides (37 in all), the final combined data set contained 12,677 redundant phosphopeptides with an estimated 0.3% false discovery rate. This final list is found in Supplemental Table 1 with hyperlinks to visualize MS/MS spectra using computer-assisted validation.

Each SEQUEST-identified phosphorylation site from all peptides in Supplemental Table 2.1 was submitted to the Ascore algorithm for improved site localization [28]. Sites with Ascore values of >19 were considered localized with near certainty while those with scores between 13 and 19 were considered localized with high certainty.

Gene Ontology (GO) annotations

All gene ontology data were analyzed with the GoMiner program[29] (evidence code set to level 3, http://discover.nci.nih.gov/gominer/) and were used as a resource for looking at the biological processes and cellular localization of the identified phosphoproteins (1159/1194 proteins were annotated with at least one GO category). The gene ontology definitions used by the GoMiner program are those annotated by the Gene Ontology Consortium (http://www.geneontology.org/).

Motif analysis

General motif classes/sequence categories were assigned given the rules previously defined [13]. Specific motifs were extracted from the data set using the Motif-X algorithm[30] (http://motif-x.med.harvard.edu). The Sanger Institute *S. pombe* database was used as the background[4] (uploaded into Motif-X in FASTA format). Candidate sequences were centered at the phosphorylated residue and extended 6 residues on each side, giving a total length of 13 amino acids for each phosphorylation site. Only extendible sites (non N/C-terminal peptides) with an Ascore >19 were used for motif extraction. The minimum reported number of occurrences for a given motif was set at 2% of the total number of phosphorylation sites found for a given residue (for example 1562 serine sites * 0.02 = 31). Only motifs with a motif score of >6 (binomial probability $<10^{-6}$) were reported. Sequence logos were automatically generated by the Weblogo program[31] (http://weblogo.berkeley.edu).

EVIN analysis for homology

The EVIN (EVolution INdexing) program http://gygi.med.harvard.edu/evin) was used to analyze conservation of phosphorylated amino acids, using multiple species alignments (currently generated by the MUSCLE program[32]). Candidate sequences were again centered at the phosphorylated residue and extended 6 residues on each side, giving a total length of 13 amino acids for each phosphorylation site, though the full protein sequence is used for alignments once the 13mer was matched to a protein sequence. In this analysis only sites with an Ascore >19 were used, except where otherwise noted. N/C terminal peptides were used in this analysis. The EVIN algorithm utilized the AL2CO program to score the amino acid frequency of all amino acids at the position of the phosphorylated residue, thus indicating the number of protein sequences in which the residue is conserved[33]. The analysis used 16 bacteria species, 21 single-cellular eukaryote species, and 14 multi-cellular eukaryote species in the alignments. The EVIN program uses the OrthoMCL data base for displaying alignments[34, 35]. Details on

the EVIN program are discussed by Ballif *et al* [36]. All data for the EVIN alignments are given under the "EVIN" tab in Supplemental Table 2.1.

Results and Discussion

Collection of the Large-Scale Phosphorylation Data Set for Fission Yeast

In order to identify a large number of phosphorylation sites, whole cell fission yeast lysate was fractionated and enriched across multiple dimensions (Figure 2.1). Six milligrams of protein from thiabendazole-treated S*chizosaccharomyces pombe* were separated by SDS-PAGE. Twelve regions of the gel were excised and proteolyzed with trypsin. The resulting peptides were split equally for enrichment of phosphopeptides by either Fe(III)-IMAC resin[21] or by titanium dioxide chromatography[22] (TiO_2). Each resulting sample was then analyzed by LC-MS/MS analysis using an LTQ-Orbitrap hybrid mass spectrometer. High mass accuracy precursor ion spectra were collected on the Orbitrap, while MS/MS spectra were generated in the linear ion trap by collision-induced dissociation (CID). In addition to the IMAC and TiO_2 data sets, a control data set was collected, consisting of peptides not retained by the IMAC enrichment (IMAC flow thru). MS/MS spectra were searched against a composite database of all *S. pombe* ORFs in the forward and reverse orientations [27], using posttranslational modification parameters for Ser, Thr and Tyr phosphorylation, methionine oxidation, and cysteine carboxyamidomethylation. Using the decoy matches as a guide, optimized thresholds were applied such that the final data sets each had a false discovery rate (FDR) <0.5% once those decoy hits were removed. The final data sets contained 5,997, 6,680, and 14 phosphopeptides from IMAC, TiO_2, and the IMAC flow through samples, respectively. In total, 2,887 unique sites were identified from 1194 proteins (Supplemental Table 2.1).

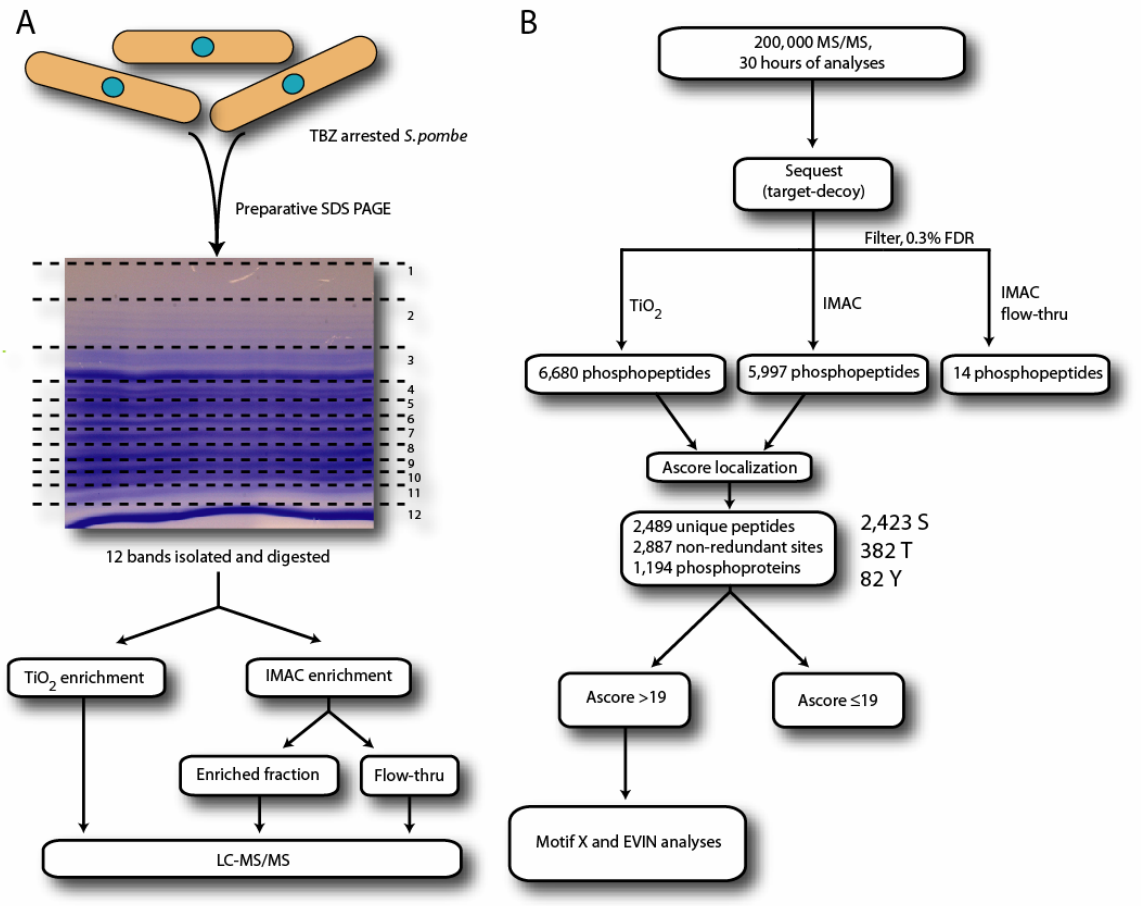


Figure 2.1. Scheme for the large-scale identification and characterization of phosphorylation sites from *S. pombe*. (A) Sample preparation and mass spectrometry. *S. pombe* was treated with thiabendazole prior to lysis to increase mitotic phosphorylation. Approximately six milligrams of total protein was separated by preparative SDS-PAGE. Twelve gel regions were excised and subjected to an in-gel digestion with trypsin. The resultant peptides were split and enriched for phosphorylation using either IMAC or TiO_2. Phosphopeptides were analyzed by reverse phased LC-MS/MS. As a control, the flow-through from the IMAC isolation was also analyzed. (B) Data processing. MS/MS spectra were searched using the SEQUEST algorithm and the target-decoy database strategy against an *S. pombe* protein database. Using decoy matches as a guide, phosphopeptide matches were filtered such that the final list contained 0.3% estimated false positives. The three data sets contained 6,680, 5,997 and 14 phosphopeptides from the TiO_2, IMAC, and IMAC flow-thru samples, respectively. The Ascore algorithm [28] was used to assign a probability of correct site localization to every site in each data set. Combining the TiO_2 and IMAC data sets resulted in the detection of 2,887 different phosphorylation sites from 2489 phosphopeptides (1,194 proteins). Sites with Ascore values >19 ($P<0.01$) were analyzed by the Motif-X [30] and EVIN[36] algorithms for motif extraction and homology analysis, respectively.

Neutral loss (NL) of phosphoric acid is commonplace in ion trap CID spectra and reduces the amount of observed backbone fragmentation (signal from b- and y- ions), limiting the proper identification of phosphoproteins [15, 37, 38]. In a representative sample from this study, almost 60% of the

spectra showed prominent neutral loss (>90% relative intensity, data not shown). The linear ion trap, however, can still produce spectra with sufficient backbone fragmentation, due to its large capacity (Figure 2.2A). The result is greatly increased signal to noise for most peaks despite signal suppression by neutral loss; consequently, many phosphopeptides can be confidently identified without collecting higher order spectra (e.g. MS^3) [15]. For example, fragmentation between peptide bonds of amino acids adjacent to the phosphoserine in Figure 2.2A localized the site to the designated serine using the Ascore algorithm [28]. Over 60% of detected phosphorylation sites were localized with near certainty (Figure 2.2B). Sites with Ascore values >19 ($P < 0.01$) were considered localized with near certainty, and those with values between 13 and19 were considered localized with high certainty ($P < 0.05$-0.01).

Fractionation by preparative SDS-PAGE allowed us to identify thousands of phosphopeptides. We previously used this approach using protein from alpha-factor arrested *S. cerevisiae* [14]. In general, many hundreds to a few thousand phosphopeptides were identified from each band (Figure 2.2C). In addition, most phosphopeptides contained only one phosphate group (Figure 2.2D). The data set was comprised of approximately 86% single phosphorylated peptides, 12% doubly phosphorylated peptides, and 2% triply phosphorylated peptides. As noted in similar experiments with *Saccharomyces cerevisiae* and in HeLa cells utilizing preparative SDS-PAGE prior to enrichment, more phosphopeptides were identified from higher molecular weight portions of the gel than the lower molecular weight portions [14, 15, 28].

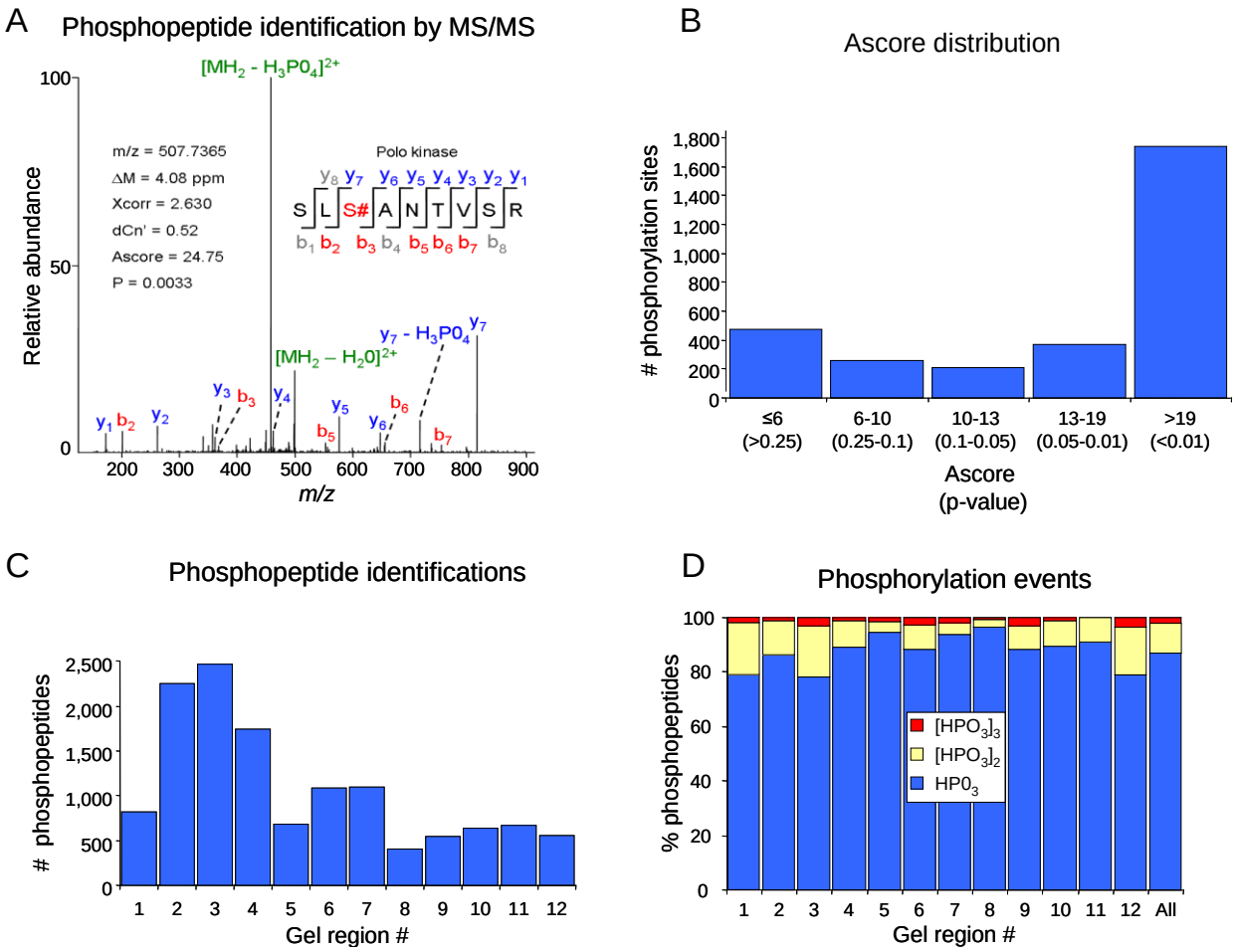


Figure 2.2. General features of the large-scale phosphorylation data set. (A) Example of an MS/MS spectrum obtained by a hybrid LTQ-Orbitrap, assigned to a peptide from Polo kinase containing phosphoserine. The precursor ion (m/z = 507.7365) was isolated and fragmented by collision-induced dissociation (CID) in the linear ion trap. Fragment ions containing the N- (*b-type* ions) or C- (*y-type* ions) terminus are labeled. Despite the very prominent neutral loss of phosphoric acid from the intact peptide ion, sufficient fragmentation at peptide bonds occurred to conclusively identify the peptide sequence and localize the site of phosphorylation from 4 different possibilities. (B) Distribution of Ascore values for the combined IMAC and TiO_2 data sets (N = 2,887 sites). The majority (60%) of all sites were considered localized with near certainty (P<0.01; Ascore > 19), while almost 75 % were localized with high certainty of greater (P<0.05; Ascore >13). (C) Distribution of all phosphopeptides by gel band. Higher molecular weight bands (>75 KDa) generally produced more phosphopeptide identifications. (D) Phosphopeptide distribution showing the number of phosphates detected on each peptide. The vast majority of phosphopeptides (86%) identified by IMAC and TiO_2 contained only one phosphate.

Evaluation of Phosphopeptide Enrichment Strategies

Since IMAC uses metal affinity interactions for phosphopeptide enrichment, and TiO_2 relies on acid-base interactions for resin competition[21, 22], the potential exists for differential enrichments

between the two methods. Overall, no large differences were detected in the number of peptides from each gel band between the IMAC and TiO_2 runs (Figure 2.3A). It is important to note however, that differences within a single band may be due to sample handling or the stochastic nature of data-dependent "shotgun" analyses. For example, very few phosphopeptides were detected in band 12 by IMAC, but >500 were identified by the TiO_2 method. Some sample to sample variability attributed to shotgun sequencing can be removed by acquiring replicates [39, 40]. Despite beginning with ~6 mg of starting material however, following IMAC and TiO_2 enrichment there was only sufficient peptide amounts for one analysis. For this reason, the twelve paired analyses were considered together, suggesting that no significant difference existed in average detected phosphopeptide numbers from either strategy ($P > 0.28$).

In contrast, the enrichment efficiency (fraction of phosphopeptides/total peptides) was statistically different between the two methods ($P < 0.002$, Figure 2.3B) with IMAC-enriched samples containing more non-phosphorylated peptides ($P < 0.007$, Supplemental Fig 1). The standard deviation for TiO_2-enriched samples was half that of IMAC (11 vs. 22%), suggesting greater consistency. Taken together, these results suggest that IMAC as applied here was less selective than TiO_2 for phosphopeptides. Finally, the TiO_2 method could likely be optimized and increase the total number of phosphopeptides isolated, perhaps by increasing the amount of resin used.

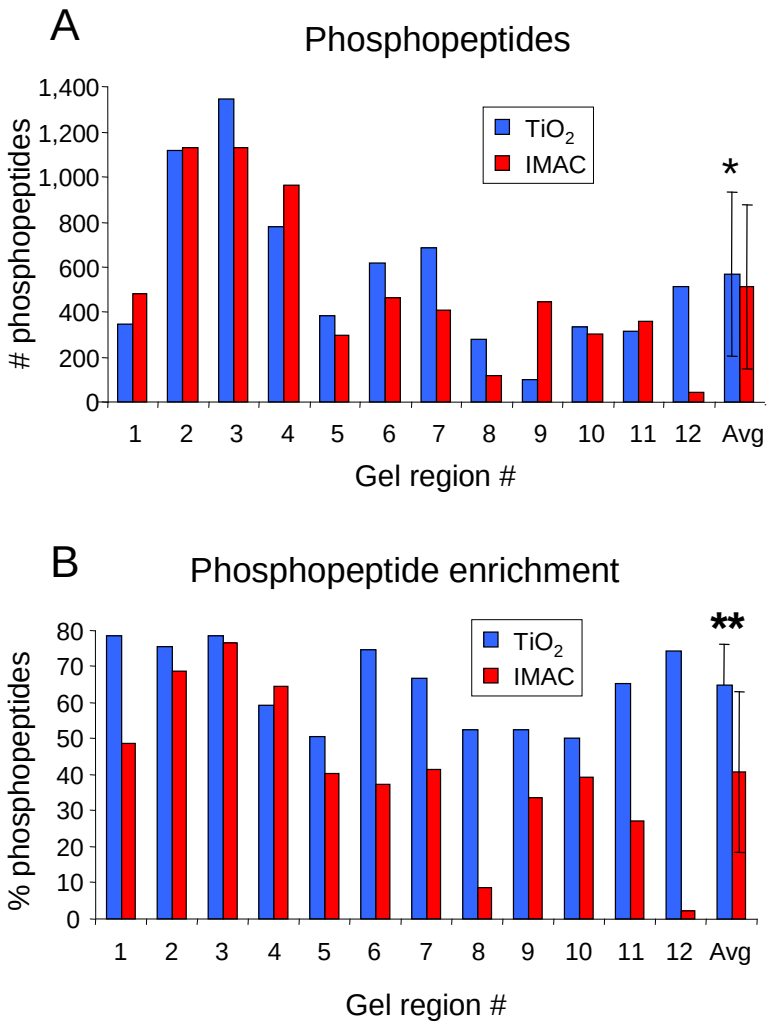


Figure 2.3. Evaluation of phosphopeptide enrichment by IMAC and TiO_2. (A) Distribution of total phosphopeptides across 12 gel bands. Although variability was seen in bands 8, 9 and 12, IMAC and TiO_2 generally gave a similar number of identification phosphopeptides within the same gel band. The average (+/- standard deviation) phosphopeptides identified by TiO_2 was 437 +/- 259 and by IMAC was 388 +/- 287. *P> 0.28; Student's paired t-Test. (B) Phosphopeptide enrichment across 12 gel bands. Enrichment was calculated as the percent of phosphopeptides in a given sample. The TiO_2 method significantly out-performed the IMAC method with respect to the percent phosphopeptides observed in a particular band. The average (+/- standard deviation) enrichment efficiency of TiO_2 was 65 ± 11%, while the IMAC average enrichment efficiency was 41 ± 22%. **P<0.002; Student's paired t-Test.

The TiO_2 enrichment method used here relies on competition between DHB and peptides for the resin. While it has been suggested that this competition is more important for inhibiting non-phosphorylated peptide binding [25], the acid-base properties associated with the TiO_2 method may enhance its selectivity for particular types of phosphopeptides compared to IMAC. To this end, a

comparison of the abundance of acidic (D, E, pT/pS/pY) and basic residues (K, R, H) found within phosphopeptide sequences by each enrichment method (Figure 2.4A and 4B) indicates no clear differences in acidic residues (analyzed per lane, P >0.76, data not shown), as both methods preferred highly acidic peptides. However, IMAC had better selectivity with >1 basic residues per peptide (analyzed per lane, $P < 10^{-9}$, data not shown). These results may be important to consider when selecting a method for phosphopeptide enrichment, as properties such as charge distribution and basic amino acid frequency may influence MS/MS performance with both CID and electron transfer dissociation (ETD), for example[38, 41]. There were, however, no significant differences in the distributions of general classes of sites (e.g., basophilic, acidophilic, etc.) (P = 1.0, Supplemental Figure 2.2). This result indicates that although the results presented in Figure 2.4B are of technical importance and affect the composition of the data set obtained, the differences in peptide preference may not substantially influence the biological conclusions that can be drawn from an experiment.

An average of 29 ± 7% of phosphorylation sites from each gel band was identified by both enrichment methods (Fig 4C). Notably, the small degree of overlap suggests that both methods can be combined to produce a much larger data set. The complementary nature of these two methods has been examined recently by other labs [23, 42]. The findings, however, may not be solely due to the different enrichment methods: the use of simple replicates of the same sample also provides large increases (~30-40% for a single replicate) in phosphopeptide identifications [40]. Whether or not this complementarity is improved by different enrichment methods vs. replicate analyses remains an unanswered question, still requiring further experimentation.

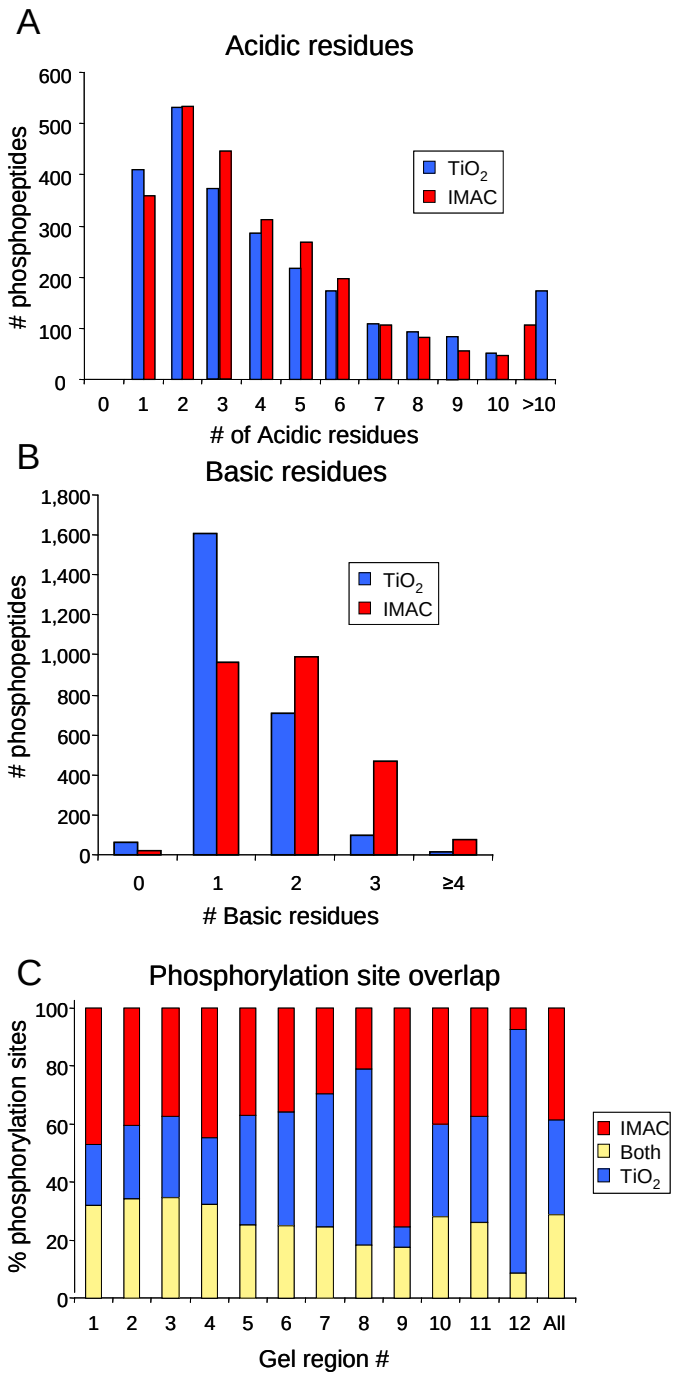


Figure 2.4. Comparison of the properties of TiO_2 and IMAC data sets. Distributions of phosphopeptides with respect to (A) acidic residues (D, E, and pS/pT/pY) and (B) basic residues (K,R, and H) in the IMAC and TiO_2 data sets. (C) Overlap in phosphorylation sites between the TiO_2 and IMAC data sets in each gel band. Though there was large variation in overlap between lanes, generally the overlap between data sets was small, averaging 29 ± 7% standard deviation. Bands 9 and 12 showed obvious deviation from the other bands due to very small overall phosphopeptide numbers in the TiO_2 and IMAC enrichments, respectively.

Motif Analysis

Observation of residue-specific motifs can reveal kinase-substrate relationships [43]. We used the Motif-X algorithm [30] to extract significant motifs from our data set (Figure 2.5). Similar to other studies [13, 14], we discovered several examples of acidic Casein Kinase II- (CKII) like motifs including sDxD, sDxE, sxED and sExE (Figure 2.5A). These motifs were found 365 times in the data set on many known substrates of CKII, including translation initiation and elongation factor subunits (eIFs and eEFs). Basophilic motifs such as the general Protein Kinase A/C (PKA/C) motif, K/Rxxs, were also abundant (Figure 2.5B). Known targets of PKA and PKC containing the phosphorylated motif were identified including the Src homology protein Cyk3 and the enzyme acetyl-CoA carboxylase. The motif for growth associated histone H1 kinase, sPK/sPR, was observed nearly 100 times in this data set (Figure 2.5C). Of specific interest is the fact that this motif was found to be phosphorylated in 10% of all occurrences of the sequence within the *S. pombe* genome (96 phosphorylated/939 total, see Supplemental Table 2.2), suggesting this motif may constitute a highly phosphorylated sequence. The protein, shugoshin, an important protein for proper chromosome segregation[44, 45] is an example of a protein containing this motif, suggesting it may be a target of histone H1 kinase activity. The common proline-directed phosphorylation motifs of sP and tP were found over 400 times in this data set (Figure 2.5D) and were the most abundant motifs detected. Notably, motifs with alterations, extensions, or combinations of known ones were identified, as in a previous report from our lab[13]. For example the tPP and RxxsP motifs appeared to be extensions of the proline-directed motif (Figure 2.5E). The mitotic kinase Cds1, transcription factors (e.g. fork-head transcription factor Fkh2), as well as many unannotated proteins contained these motifs. These motifs may constitute entirely new kinase specificities, or perhaps they modulate known kinase activities by bridging signaling pathways, able to accept phosphorylation signals from multiple kinases.

Despite small numbers of sites, one tyrosine motif, yxxxRxY was significantly represented in this study (Figure 2.5F). This motif had been previously identified in a large-scale phosphorylation study in mammals[13], but has not yet been assigned to a specific kinase activity. This motif was observed here in certain kinases, including Gsk3 and the MAP kinase Sty1. While there are no known tyrosine kinases in yeast, there are dual specificity kinases. This pattern may represent the tyrosine kinase activity of a dual-specificity kinase. Finally, a novel tG motif was observed (Figure 2.5F), though its function is currently unknown.

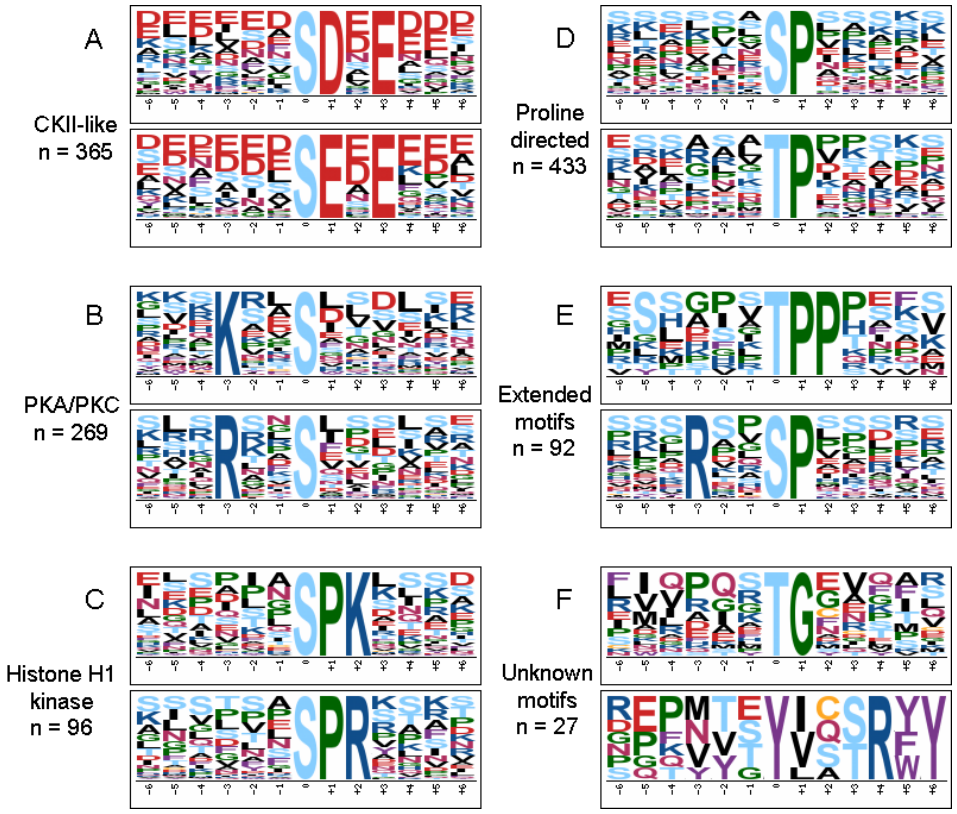


Figure 2.5. Phosphorylation motifs extracted using the Motif-X algorithm. Only sites with an Ascore >19, excluding N- and C- terminal peptides, were considered in this analysis (N=1,715). Sites were extended by 6 residues on each side of a given phosphorylated residue. The complete list of all returned motifs is given in Supplemental Table 2. (A) Acidic motifs showing CKII-like activity; these motifs were observed 365 instances. (B) Basic motifs, substrates for PKA/PKC, were found 269 times. (C) Growth associated histone H1 kinase activity is often ascribed to these motifs, which were found 96 times in the data set. (D) Proline-directed motifs, common to mitotically phosphorylated proteins, were detected more than 400 times. (E) Uncharacterized motifs that appeared to be extensions of known motifs. tPP appeared only 12 times although it was 28 fold overrepresented compared to the background and seemed to be an extension of a proline-directed motif. RxxsP and RxxtP motifs were found 80 times and appeared to be a combination of both basophilic and proline-directed motifs. (F) Uncharacterized motifs with no known similarity. Sequence specific kinase motifs were annotated by the Human Protein Reference Database[59].

Correlation of Phosphorylation Data with Known Biological Pathways

One goal of proteomics is to gain insight into complex biological processes through the creation of high quality databases. In this study, a large data set of phosphorylation sites was assembled for several reasons. First, it is a resource for researchers interested in phosphorylation-dependent regulation of all biological processes. In total, almost 25% of known *S. pombe* ORFs were found with at least one phosphorylation event, and GO analysis revealed that these proteins span multiple categories for biological process and cellular localization (Supplemental Figure 2.3). Second, this database is enriched for mitotic phosphorylation and should prove particularly useful for cell-cycle related processes. Additional GO analysis combined with recent literature were useful for finding proteins involved in mitotic commitment (G_2/M transition) and regulation of the spindle assembly checkpoint (SAC)/Chromosome segregation; moreover, the SAC GO category was over represented two-fold in this data set ($P< 0.01$, one-sided Fisher exact test[29]) . Finally, because many *S. pombe* proteins have human homologs, these data promote a rich environment to study highly-conserved phosphorylation events.

Much of the cell cycle is controlled by phosphorylation. For example, three such pathways involved in entry and progression through M-phase which are regulated by phosphorylation include Cdc2 activation, SAC regulation, and inhibition of the APC. Identification of novel phosphorylation sites on known regulatory proteins may suggest new roles of signaling control. For example, new sites were detected in Cds1, Rad24 and Rad25. Cds1 is known to regulate Cdc25 phosphorylation[46, 47], while Rad24/25 regulate the nuclear export of phosphorylated Cdc25[48]. If these novel sites affect the molecular function of these proteins, they very well may play a role in mitotic commitment. The spindle pole protein Cut12 and Polo kinase, which have been shown to interact in regulatory manner[49], both showed novel phosphorylation in this study. Although one cannot specifically link these new phosphorylation events to direct regulation, the potential modulation of this very important kinase activity remains intriguing. In a similar manner the SAC kinase Bub1, a major player in the mitotic arrest

through its maintenance of Mad2 and Mad3 activity[50, 51], showed novel phosphorylation. Bub1 may also directly phosphorylate and inhibit Cdc20, further contributing to the metaphase checkpoint[52]. Finally, regulators of APC activity, including Shugoshin, Survivin/Bir1, and INCENP-like/Pic1 proteins showed novel phosphorylation in this study. These proteins are known to influence the localization and activity of aurora kinase, which has been shown to be one of the most important proteins for transitioning into anaphase[44, 45, 53, 54]. Many of the sites mentioned here contain a proline directed motif, which is a common motif seen in Cdk mediated cell cycle control, supporting the potential for the new sites to be involved in mitotic regulation. These data are summarized in Figure 2.6.

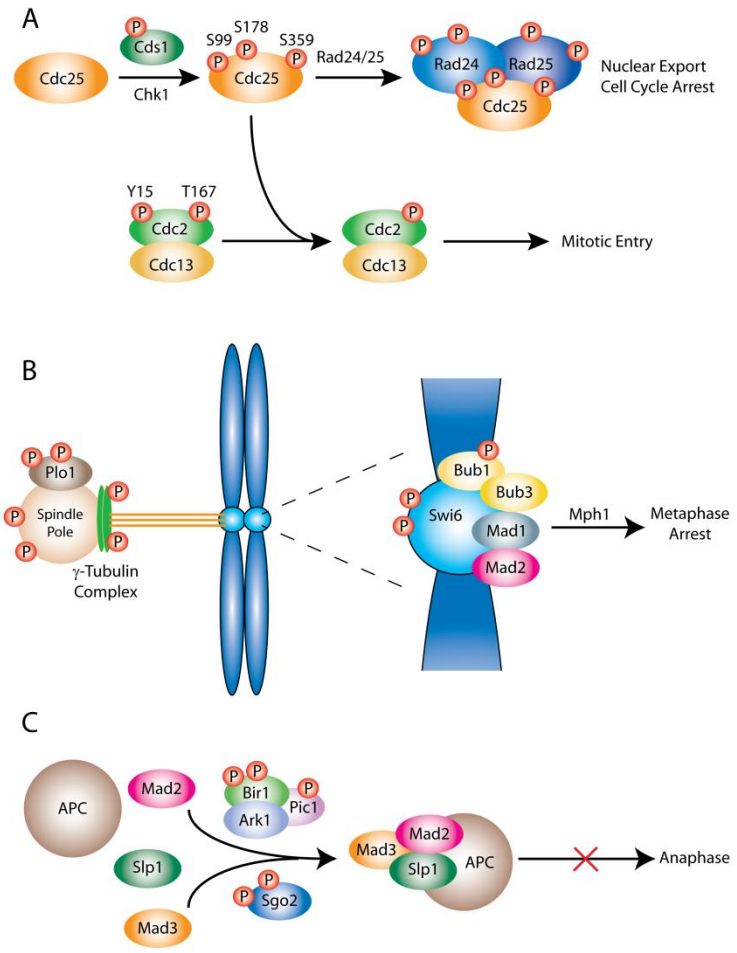


Figure 2.6. Assignment of phosphorylation data to biologically relevant pathways involved in entry and progression through M-phase. Proteins shown with one phosphate group were found to be singly phosphorylated, and those showing two phosphate groups were found to be multiply phosphorylated. (A) The control of mitotic entry by Cdc25. Cdc25 dephosphorylates Y15 on Cdc2, removing this inhibitory signal added by the Wee1 kinase. T167 on Cdc2 is known to be an activating phosphorylation event[60]. The kinase Cds1 phosphorylates Cdc25 at several positions including S99 and S359[46, 47] (S178 phosphorylation on Cdc25 is an inhibitory signal put on by Chk1 kinase[61]), making it available for nuclear export by Rad24/25[48]. Novel phosphorylation sites were found on Cds1, Rad24, and Rad25, with several being identified on the last 2 proteins, suggesting there may be additional factors controlling the export of and Cdc25/M-phase commitment (see Supplemental Table 2 for specific sites). (B). The spindle assembly checkpoint (SAC). Proteins annotated with gene ontology (GO) categories for the SAC and chromosome segregation, as well as mitotically relevant proteins found in the literature were considered. Many new phosphorylation sites were discovered on the spindle pole (Cut12) and associated proteins, including the gamma-tubulin (Alp4/6) complex and polo kinase, all of which play roles in spindle assembly[49, 62]. Centromere associated proteins including Swi6 and Bub1, known contributors to the activation of the SAC[50, 63], were found to be phosphorylated at novel sites as well. It is believed that the SAC signal may be transduced by the kinase Mph1[64]. (C) The inhibition of the anaphase promoting complex (APC), adapted from[64]. Aurora kinase plays a role in promoting the inhibition of the APC, and shugoshin (Sgo2), survivin (Bir1) and INCENP-like protein (Pic1) promote the assembly and enhance the activity of aurora kinase at the centromere[44, 45, 53, 54], all of which showed novel phosphorylation sites. These new sites provide starting points for future studies and may modulate the progression through mitosis by regulating the key steps.

Conservation of Phosphorylation Sites Across Species

Lipin (also called Ned1) is an example of a protein found to be highly phosphorylated in this study, and has a mammalian homolog also known to be a phosphoprotein (figure 2.7). This protein is implicated in mammalian metabolic diseases such as lipodystrophy and insulin resistance [55]. In one study, mutations of *S. pombe* lipin caused high levels of aberrant chromosomal segregation with oddly shaped nuclei and were shown to cause hypersensitivity to microtubule depolymerizing agents [56]. This same study also found that Ned1 was heavily phosphorylated (at unspecified sites) in a growth-related manner, implicating it in the relationship between growth and division. Both yeast and mammalian forms of lipin contain two domains, a highly conserved N-terminal domain and a C-terminal domain.

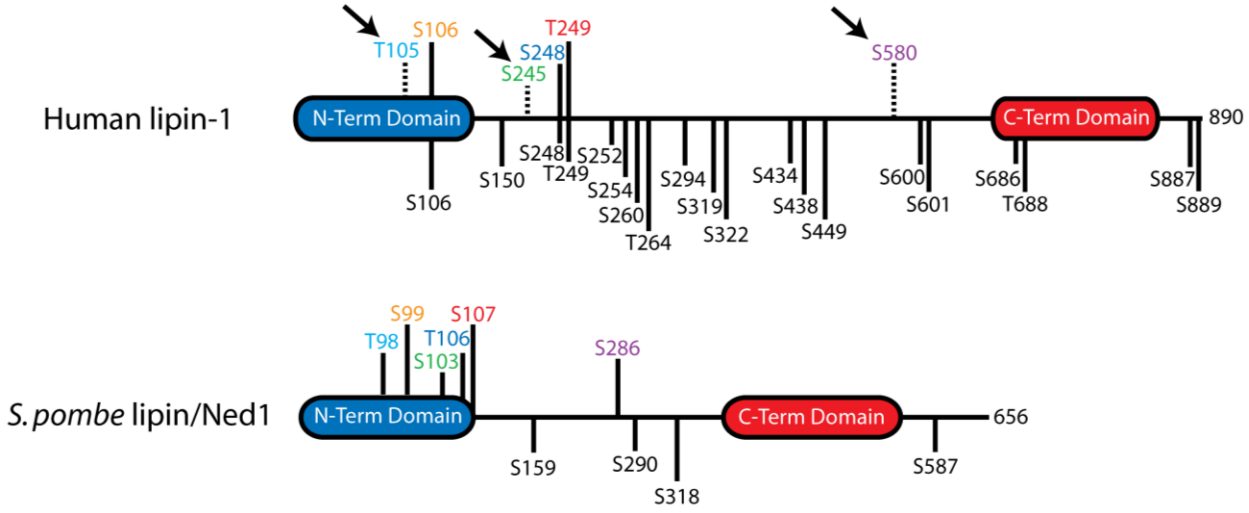


Figure 2.7. Phosphorylation sites in *S. pombe* often validate and can even predict conserved phosphorylation in higher eukaryotes. The protein lipin has homologs in both *S. pombe* (lipin/Ned1) and humans (lipin-1). A recent study[57] showed that the mammalian protein lipin-1 is highly phosphorylated, and 20 sites have been annotated in the human protein specifically. Known sites are shown with a black line below human lipin-1. Although it was previously known to be a phosphoprotein[56], no sites for the *S. pombe* homolog (lipin/Ned1) were present in the literature prior to this study. We identified 10 phosphorylation sites from lipin/Ned1. Using the EVIN program[36], several of the phosphorylation sites identified in our study were found to share homology with known phosphorylation sites in human lipin-1. Color coded homologous sites in *S. pombe* lipin/Ned1 are shown as lines extending above the protein whereas non-homologous sites are shown below the protein. Interestingly, several sites were conserved phosphorylation events between the two species. Finally, three additional sites from *S. pombe* lipin (T98, S103, and S286) have high homology to three human lipin-1 residues (T105, S245, S580, shown with dashed lines and highlighted with arrows), predicting that the human sites are phosphorylated as well.

In this study, ten sites were identified from this protein, half of which were in the N-terminal domain. Only sites with an Ascore >19 (P <0.01), or an Ascore between 13 and19 (P <0.05) with manual validation were considered. In the mouse homolog of lipin, 23 phosphorylation sites were identified by lipin overexpression using 3T3-L1 adipocytes [57]. Twenty of these sites are annotated as phosphorylated in humans based on high homology in the Phosphosite database[58]. Several of these sites were found to be conserved between *S. pombe* and humans through the use of the EVIN program[36], such as S99 in yeast and S106 in humans (Supplemental Table 2.1, "EVIN" tab). Moreover, this analysis predicted three new phosphorylation sites on human lipin from high conservation to those found in *S. pombe*, such as the human T105 based upon its alignment with T98 in yeast. Most notably, no sites from any form of mammalian lipin have been discovered based on large-scale studies [13, 16, 17]; all known sites on mammalian lipin were from the single overexpression study [57]. In the whole data set, EVIN provided alignments for 55% (977/1772) of the localized phosphorylation sites, and nearly 20% (188/977) of those sites showed >40% conservation of a phosphorylatable residue, suggesting some phosphorylation events may be conserved. Because the *S. pombe* proteome is less complex than higher eukaryotes, a higher proportion of the expressed proteome can be analyzed without additional fractionation. All these data taken together suggest that this conservation approach could be useful for studying difficult or uncharacterized mammalian proteins, allowing the characterization of potentially critical biology based on ancient regulation in simpler organisms.

Conclusions

This work represents the largest yeast phosphorylation data set collected to date. In addition, these data are the only available large-scale phosphorylation site reference for fission yeast. As such they provide a resource for future experimentation. Furthermore, many of the identified sites in this

analysis may be of great interest to those studying cell-cycle regulation, as these data are enriched in mitotic phosphorylation. This analysis provides a general framework for large-scale phosphorylation analysis, combining pre-fractionation techniques with phosphopeptide enrichment strategies. A comparison of two popular phosphopeptide enrichment methods, IMAC and TiO_2, revealed that they are complementary, providing unique sets of phosphopeptides which contain unique properties. Finally, the conservation of phosphorylation sites across many species is given, and provides a framework for predicting potentially relevant phosphorylation events in higher eukaryotes.

Acknowledgements

This work was supported in part by NIH grants HG3456 and HG3616 (to S.P.G.) and a postdoctoral fellowship from the Spanish Ministry of Education and Science (to J.V.). We would like to thank Danesh Moazed for providing the *S. pombe* for culture and thiabendazole. We thank Corey Bakalarski, Joshua Elias, Sean Beausoleil, and Richard Carey for helpful bioinformatics assistance.

References

1. Russell, P.; Nurse, P., cdc25+ functions as an inducer in the mitotic control of fission yeast. *Cell* **1986,** 45, (1), 145-53.
2. Featherstone, C.; Russell, P., Fission yeast p107wee1 mitotic inhibitor is a tyrosine/serine kinase. *Nature* **1991,** 349, (6312), 808-11.
3. Den Haese, G. J.; Walworth, N.; Carr, A. M.; Gould, K. L., The Wee1 protein kinase regulates T14 phosphorylation of fission yeast Cdc2. *Mol Biol Cell* **1995,** 6, (4), 371-85.
4. Wood, V.; Gwilliam, R.; Rajandream, M. A.; Lyne, M.; Lyne, R.; Stewart, A.; Sgouros, J.; Peat, N.; Hayles, J.; Baker, S.; Basham, D.; Bowman, S.; Brooks, K.; Brown, D.; Brown, S.; Chillingworth, T.; Churcher, C.; Collins, M.; Connor, R.; Cronin, A.; Davis, P.; Feltwell, T.; Fraser, A.; Gentles, S.; Goble, A.; Hamlin, N.; Harris, D.; Hidalgo, J.; Hodgson, G.; Holroyd, S.; Hornsby, T.; Howarth, S.; Huckle, E. J.; Hunt, S.; Jagels, K.; James, K.; Jones, L.; Jones, M.; Leather, S.; McDonald, S.; McLean, J.; Mooney, P.; Moule, S.; Mungall, K.; Murphy, L.; Niblett, D.; Odell, C.; Oliver, K.; O'Neil, S.; Pearson, D.; Quail, M. A.; Rabbinowitsch, E.; Rutherford, K.; Rutter, S.; Saunders, D.; Seeger, K.; Sharp, S.; Skelton, J.; Simmonds, M.; Squares, R.; Squares, S.; Stevens, K.; Taylor, K.; Taylor, R. G.; Tivey, A.; Walsh, S.; Warren, T.; Whitehead, S.; Woodward, J.; Volckaert, G.; Aert, R.; Robben, J.; Grymonprez, B.; Weltjens, I.;

Vanstreels, E.; Rieger, M.; Schafer, M.; Muller-Auer, S.; Gabel, C.; Fuchs, M.; Dusterhoft, A.; Fritzc, C.; Holzer, E.; Moestl, D.; Hilbert, H.; Borzym, K.; Langer, I.; Beck, A.; Lehrach, H.; Reinhardt, R.; Pohl, T. M.; Eger, P.; Zimmermann, W.; Wedler, H.; Wambutt, R.; Purnelle, B.; Goffeau, A.; Cadieu, E.; Dreano, S.; Gloux, S.; Lelaure, V.; Mottier, S.; Galibert, F.; Aves, S. J.; Xiang, Z.; Hunt, C.; Moore, K.; Hurst, S. M.; Lucas, M.; Rochet, M.; Gaillardin, C.; Tallada, V. A.; Garzon, A.; Thode, G.; Daga, R. R.; Cruzado, L.; Jimenez, J.; Sanchez, M.; del Rey, F.; Benito, J.; Dominguez, A.; Revuelta, J. L.; Moreno, S.; Armstrong, J.; Forsburg, S. L.; Cerutti, L.; Lowe, T.; McCombie, W. R.; Paulsen, I.; Potashkin, J.; Shpakovski, G. V.; Ussery, D.; Barrell, B. G.; Nurse, P., The genome sequence of Schizosaccharomyces pombe. *Nature* **2002,** 415, (6874), 871-80.

5. Matsuyama, A.; Arai, R.; Yashiroda, Y.; Shirai, A.; Kamata, A.; Sekido, S.; Kobayashi, Y.; Hashimoto, A.; Hamamoto, M.; Hiraoka, Y.; Horinouchi, S.; Yoshida, M., ORFeome cloning and global analysis of protein localization in the fission yeast Schizosaccharomyces pombe. *Nat Biotechnol* **2006,** 24, (7), 841-7.

6. Sipiczki, M., Where does fission yeast sit on the tree of life? *Genome Biol* **2000,** 1, (2), REVIEWS1011.

7. Sipiczki, M., Phylogenesis of fission yeasts. Contradictions surrounding the origin of a century old genus. *Antonie Van Leeuwenhoek* **1995,** 68, (2), 119-49.

8. Lucas, J. I.; Marin, I., A new evolutionary paradigm for the Parkinson disease gene DJ-1. *Mol Biol Evol* **2007,** 24, (2), 551-61.

9. Tvegard, T.; Soltani, H.; Skjolberg, H. C.; Krohn, M.; Nilssen, E. A.; Kearsey, S. E.; Grallert, B.; Boye, E., A novel checkpoint mechanism regulating the G1/S transition. *Genes Dev* **2007,** 21, (6), 649-54.

10. Wolfe, B. A.; McDonald, W. H.; Yates, J. R., 3rd; Gould, K. L., Phospho-regulation of the Cdc14/Clp1 phosphatase delays late mitotic events in S. pombe. *Dev Cell* **2006,** 11, (3), 423-30.

11. Dove, S. K.; Cooke, F. T.; Douglas, M. R.; Sayers, L. G.; Parker, P. J.; Michell, R. H., Osmotic stress activates phosphatidylinositol-3,5-bisphosphate synthesis. *Nature* **1997,** 390, (6656), 187-92.

12. Matsuo, T.; Kubo, Y.; Watanabe, Y.; Yamamoto, M., Schizosaccharomyces pombe AGC family kinase Gad8p forms a conserved signaling module with TOR and PDK1-like kinases. *Embo J* **2003,** 22, (12), 3073-83.

13. Villen, J.; Beausoleil, S. A.; Gerber, S. A.; Gygi, S. P., Large-scale phosphorylation analysis of mouse liver. *Proc Natl Acad Sci U S A* **2007,** 104, (5), 1488-93.

14. Li, X.; Gerber, S. A.; Rudner, A. D.; Beausoleil, S. A.; Haas, W.; Villen, J.; Elias, J. E.; Gygi, S. P., Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae. *J Proteome Res* **2007,** 6, (3), 1190-7.

15. Beausoleil, S. A.; Jedrychowski, M.; Schwartz, D.; Elias, J. E.; Villen, J.; Li, J.; Cohn, M. A.; Cantley, L. C.; Gygi, S. P., Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A* **2004,** 101, (33), 12130-5.

16. Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M., Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **2006,** 127, (3), 635-48.

17. Trinidad, J. C.; Specht, C. G.; Thalhammer, A.; Schoepfer, R.; Burlingame, A. L., Comprehensive identification of phosphorylation sites in postsynaptic density preparations. *Mol Cell Proteomics* **2006,** 5, (5), 914-22.

18. Gruhler, A.; Olsen, J. V.; Mohammed, S.; Mortensen, P.; Faergeman, N. J.; Mann, M.; Jensen, O. N., Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol Cell Proteomics* **2005,** 4, (3), 310-27.

19. Rush, J.; Moritz, A.; Lee, K. A.; Guo, A.; Goss, V. L.; Spek, E. J.; Zhang, H.; Zha, X. M.; Polakiewicz, R. D.; Comb, M. J., Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat Biotechnol* **2005,** 23, (1), 94-101.

20. Matsuoka, S.; Ballif, B. A.; Smogorzewska, A.; McDonald, E. R., 3rd; Hurov, K. E.; Luo, J.; Bakalarski, C. E.; Zhao, Z.; Solimini, N.; Lerenthal, Y.; Shiloh, Y.; Gygi, S. P.; Elledge, S. J., ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* **2007,** 316, (5828), 1160-6.
21. Scanff, P.; Yvon, M.; Pelissier, J. P., Immobilized Fe3+ affinity chromatographic isolation of phosphopeptides. *J Chromatogr* **1991,** 539, (2), 425-32.
22. Pinkse, M. W.; Uitto, P. M.; Hilhorst, M. J.; Ooms, B.; Heck, A. J., Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. *Anal Chem* **2004,** 76, (14), 3935-43.
23. Bodenmiller, B.; Mueller, L. N.; Mueller, M.; Domon, B.; Aebersold, R., Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nat Methods* **2007,** 4, (3), 231-7.
24. Rappsilber, J.; Ishihama, Y.; Mann, M., Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* **2003,** 75, (3), 663-70.
25. Larsen, M. R.; Thingholm, T. E.; Jensen, O. N.; Roepstorff, P.; Jorgensen, T. J., Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol Cell Proteomics* **2005,** 4, (7), 873-86.
26. Haas, W.; Faherty, B. K.; Gerber, S. A.; Elias, J. E.; Beausoleil, S. A.; Bakalarski, C. E.; Li, X.; Villen, J.; Gygi, S. P., Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Mol Cell Proteomics* **2006,** 5, (7), 1326-37.
27. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007,** 4, (3), 207-14.
28. Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P., A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* **2006,** 24, (10), 1285-92.
29. Zeeberg, B. R.; Feng, W.; Wang, G.; Wang, M. D.; Fojo, A. T.; Sunshine, M.; Narasimhan, S.; Kane, D. W.; Reinhold, W. C.; Lababidi, S.; Bussey, K. J.; Riss, J.; Barrett, J. C.; Weinstein, J. N., GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* **2003,** 4, (4), R28.
30. Schwartz, D.; Gygi, S. P., An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol* **2005,** 23, (11), 1391-8.
31. Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E., WebLogo: a sequence logo generator. *Genome Res* **2004,** 14, (6), 1188-90.
32. Edgar, R. C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **2004,** 32, (5), 1792-7.
33. Pei, J.; Grishin, N. V., AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* **2001,** 17, (8), 700-12.
34. Li, L.; Stoeckert, C. J., Jr.; Roos, D. S., OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **2003,** 13, (9), 2178-89.
35. Chen, F.; Mackey, A. J.; Stoeckert, C. J., Jr.; Roos, D. S., OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **2006,** 34, (Database issue), D363-8.
36. Ballif, B. A.; Carey, G. R.; Sunyaev, S. R.; Gygi, S. P., Large-Scale Identification and Evolution Indexing of Tyrosine Phosphorylation Sites from Murine Brain. *J Proteome Res*, in press.
37. DeGnore, J. P.; Qin, J., Fragmentation of phosphopeptides in an ion trap mass spectrometer. *J Am Soc Mass Spectrom* **1998,** 9, (11), 1175-88.
38. Mikesh, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E.; Shabanowitz, J.; Hunt, D. F., The utility of ETD mass spectrometry in proteomic analysis. *Biochim Biophys Acta* **2006,** 1764, (12), 1811-22.

39. Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P., Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Methods* **2005,** 2, (9), 667-75.
40. Bakalarski, C. E.; Haas, W.; Dephoure, N. E.; Gygi, S. P., The effects of mass accuracy, data acquisition speed, and search algorithm choice on peptide identification rates in phosphoproteomics. *Anal Bioanal Chem* **2007**.
41. Good, D. M.; Wirtala, M.; McAlister, G. C.; Coon, J. J., Performance characteristics of electron transfer dissociation mass spectrometry. *Mol Cell Proteomics* **2007**.
42. Liang, X.; Fonnum, G.; Hajivandi, M.; Stene, T.; Kjus, N. H.; Ragnhildstveit, E.; Amshey, J. W.; Predki, P.; Pope, R. M., Quantitative Comparison of IMAC and TiO(2) Surfaces Used in the Study of Regulated, Dynamic Protein Phosphorylation. *J Am Soc Mass Spectrom* **2007**.
43. Peri, S.; Navarro, J. D.; Kristiansen, T. Z.; Amanchy, R.; Surendranath, V.; Muthusamy, B.; Gandhi, T. K.; Chandrika, K. N.; Deshpande, N.; Suresh, S.; Rashmi, B. P.; Shanker, K.; Padma, N.; Niranjan, V.; Harsha, H. C.; Talreja, N.; Vrushabendra, B. M.; Ramya, M. A.; Yatish, A. J.; Joy, M.; Shivashankar, H. N.; Kavitha, M. P.; Menezes, M.; Choudhury, D. R.; Ghosh, N.; Saravana, R.; Chandran, S.; Mohan, S.; Jonnalagadda, C. K.; Prasad, C. K.; Kumar-Sinha, C.; Deshpande, K. S.; Pandey, A., Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res* **2004,** 32, (Database issue), D497-501.
44. Kawashima, S. A.; Tsukahara, T.; Langegger, M.; Hauf, S.; Kitajima, T. S.; Watanabe, Y., Shugoshin enables tension-generating attachment of kinetochores by loading Aurora to centromeres. *Genes Dev* **2007,** 21, (4), 420-35.
45. Vanoosthuyse, V.; Prykhozhij, S.; Hardwick, K. G., Shugoshin 2 regulates localization of the chromosomal passenger proteins in fission yeast mitosis. *Mol Biol Cell* **2007,** 18, (5), 1657-69.
46. Zeng, Y.; Forbes, K. C.; Wu, Z.; Moreno, S.; Piwnica-Worms, H.; Enoch, T., Replication checkpoint requires phosphorylation of the phosphatase Cdc25 by Cds1 or Chk1. *Nature* **1998,** 395, (6701), 507-10.
47. Furnari, B.; Blasina, A.; Boddy, M. N.; McGowan, C. H.; Russell, P., Cdc25 inhibited in vivo and in vitro by checkpoint kinases Cds1 and Chk1. *Mol Biol Cell* **1999,** 10, (4), 833-45.
48. Zeng, Y.; Piwnica-Worms, H., DNA damage and replication checkpoints in fission yeast require nuclear exclusion of the Cdc25 phosphatase via 14-3-3 binding. *Mol Cell Biol* **1999,** 19, (11), 7410-9.
49. MacIver, F. H.; Tanaka, K.; Robertson, A. M.; Hagan, I. M., Physical and functional interactions between polo kinase and the spindle pole component Cut12 regulate mitotic commitment in S. pombe. *Genes Dev* **2003,** 17, (12), 1507-23.
50. Bernard, P.; Hardwick, K.; Javerzat, J. P., Fission yeast bub1 is a mitotic centromere protein essential for the spindle checkpoint and the preservation of correct ploidy through mitosis. *J Cell Biol* **1998,** 143, (7), 1775-87.
51. Morrow, C. J.; Tighe, A.; Johnson, V. L.; Scott, M. I.; Ditchfield, C.; Taylor, S. S., Bub1 and aurora B cooperate to maintain BubR1-mediated inhibition of APC/CCdc20. *J Cell Sci* **2005,** 118, (Pt 16), 3639-52.
52. Tang, Z.; Shu, H.; Oncel, D.; Chen, S.; Yu, H., Phosphorylation of Cdc20 by Bub1 provides a catalytic mechanism for APC/C inhibition by the spindle checkpoint. *Mol Cell* **2004,** 16, (3), 387-97.
53. Leverson, J. D.; Huang, H. K.; Forsburg, S. L.; Hunter, T., The Schizosaccharomyces pombe aurora-related kinase Ark1 interacts with the inner centromere protein Pic1 and mediates chromosome segregation and cytokinesis. *Mol Biol Cell* **2002,** 13, (4), 1132-43.
54. Petersen, J.; Hagan, I. M., S. pombe aurora kinase/survivin is required for chromosome condensation and the spindle checkpoint attachment response. *Curr Biol* **2003,** 13, (7), 590-7.
55. Reitman, M. L., The fat and thin of lipin. *Cell Metab* **2005,** 1, (1), 5-6.
56. Tange, Y.; Hirata, A.; Niwa, O., An evolutionarily conserved fission yeast protein, Ned1, implicated in normal nuclear morphology and chromosome stability, interacts with Dis3, Pim1/RCC1 and an essential nucleoporin. *J Cell Sci* **2002,** 115, (Pt 22), 4375-85.

57. Harris, T. E.; Huffman, T. A.; Chi, A.; Shabanowitz, J.; Hunt, D. F.; Kumar, A.; Lawrence, J. C., Jr., Insulin controls subcellular localization and multisite phosphorylation of the phosphatidic acid phosphatase, lipin 1. *J Biol Chem* **2007,** 282, (1), 277-86.
58. Hornbeck, P. V.; Chabra, I.; Kornhauser, J. M.; Skrzypek, E.; Zhang, B., PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* **2004,** 4, (6), 1551-61.
59. Amanchy, R.; Periaswamy, B.; Mathivanan, S.; Reddy, R.; Tattikota, S. G.; Pandey, A., A curated compendium of phosphorylation motifs. *Nat Biotechnol* **2007,** 25, (3), 285-6.
60. Gould, K. L.; Moreno, S.; Owen, D. J.; Sazer, S.; Nurse, P., Phosphorylation at Thr167 is required for Schizosaccharomyces pombe p34cdc2 function. *Embo J* **1991,** 10, (11), 3297-309.
61. Chen, M. S.; Ryan, C. E.; Piwnica-Worms, H., Chk1 kinase negatively regulates mitotic function of Cdc25A phosphatase through 14-3-3 binding. *Mol Cell Biol* **2003,** 23, (21), 7488-97.
62. Vardy, L.; Toda, T., The fission yeast gamma-tubulin complex is required in G(1) phase and is a component of the spindle assembly checkpoint. *Embo J* **2000,** 19, (22), 6098-111.
63. Bernard, P.; Maure, J. F.; Partridge, J. F.; Genier, S.; Javerzat, J. P.; Allshire, R. C., Requirement of heterochromatin for cohesion at centromeres. *Science* **2001,** 294, (5551), 2539-42.
64. Millband, D. N.; Hardwick, K. G., Fission yeast Mad3p is required for Mad2p to inhibit the anaphase-promoting complex and localizes to kinetochores in a Bub1p-, Bub3p-, and Mph1p-dependent manner. *Mol Cell Biol* **2002,** 22, (8), 2728-42.

Chapter 3

Quantitative Comparison of the Fasted and Re-fed Mouse Liver Phosphoproteomes Using Lower pH Reductive Dimethylation

Attributions:

- This chapter contains work which is currently accepted for publication as Wilson-Grady, J.T., Haas, W., and Gygi, S.P. Quantitative Comparison of the Fasted and Re-fed Mouse Liver Phosphoproteomes Using Lower pH Reductive Dimethylation. *Methods* (2013).

- J.T. Wilson-Grady performed the experimental steps and data analysis, as well as manuscript preparation. All bioinformatic processing was performed using in-house software developed by the "GFY Development Team."

- W. Haas contributed key guidance and was instrumental in discovering the pH dependence of the dimethylation reaction, as well as critiqued the manuscript.

- S.P. Gygi advised the project.

Abstract

Phosphorylation is a common but crucial protein posttranslational modification occurring in virtually all known species. A successful technique for identifying phosphorylation sites is via liquid chromatography-tandem mass spectrometry (LC-MS/MS). In addition to identification, the introduction of stable isotopes allows for LC-MS based quantification of thousands of phosphorylation sites. Historically, stable isotope labeling by amino acids in cell culture (SILAC) has been the preferred method for introducing stable isotopes for quantification. SILAC is not well suited, however, for quantitative proteomics in larger animals. The introduction of stable isotopes instead by reductive dimethylation is an alternative for performing quantitative proteomics in animal tissues. Here we present an improved reductive dimethylation protocol and demonstrate the application of this method in the analysis of the fasted vs. re-fed mouse liver phosphoproteome. In our analysis, greater than 8500 sites were identified from ~2700 phosphoproteins. Nearly 7400 phosphorylation events from ~2300 phosphoproteins were reliably quantified. Using a 2-fold change as a cutoff, 390 phosphorylation sites were found to change between fasted and re-fed mice, many of which may have interesting biological interpretations.

Introduction

Phosphorylation is a key mediator of virtually every cellular process. Indeed, many proteins are phosphorylated on multiple residues simultaneously [1], thus demonstrating its complex nature. Phosphorylation analysis is a large field and encompasses efforts to understand phosphorylation dynamics and how they affect specific biological processes. These processes span from regulating protein-protein interactions to orchestrating complex signal transduction cascades involved in human diseases. For these reasons, new technology that serves to increase the identification and/or quantification of phosphorylation sites is valuable to the scientific community.

The preferred method for large-scale identification and quantification of phosphorylation sites is liquid chromatography-tandem mass spectrometry (LC-MS/MS). Due to the low abundance of phosphorylated species, orthogonal separation techniques and phosphopeptide enrichment methods have vastly improved the ability to detect phosphopeptides [2-4]. Studies relying heavily on LC-MS/MS and phosphopeptide enrichment have been able to identify thousands of phosphorylation sites in a variety of organisms, tissues, and cell lines [1, 5-9]. Quantifying phosphorylation sites has been accomplished commonly through the use of stable isotope labeling by amino acids in cell culture (SILAC) [10, 11]. Though the use of SILAC has been valuable to the field, it has several limitations as a result of the need to metabolically label the cells or tissue. The introduction of stable isotopes by the use of reductive dimethylation overcomes this limitation and is applicable to large scale proteomics[12].

Reductive dimethylation allows one to perform tissue-based quantitative proteomics/phosphoproteomics with several advantages over other available methods [8]. First the cost of labeling > 10 mg of peptides is less than one dollar, compared to several thousand dollars for SILAC tissue samples. Unlike SILAC, reductive dimethylation can directly compare paired samples, such as mouse littermates or cancerous vs. normal tissue from the same individual.

Despite all the advantages of reductive dimethylation, few labs have taken advantage of this technology, especially in the context of phosphorylation, (for example [13, 14]). Here we present our method for tissue-labeling with reductive dimethylation, for phosphorylation analysis. We examined the reaction pH and applied our pipeline to the large-scale analysis of the phosphorylation state in livers from fasted vs. re-fed mice.

Rationale

Through experimentation with reductive dimethylation, we noticed that these datasets contained fewer peptide matches compared to unlabeled samples, in terms of total identifications and the percent of matched spectra (data not shown). We evaluated pH as a factor affecting the identification rate. The motivation for this publication is to highlight the use of a lower pH for the reductive dimethylation reaction, and how it can be successfully applied to the quantification of phosphorylation changes in the example of fasted vs. re-fed mouse liver samples.

Materials and Methods

The dimethylation reaction chemistry and the workflow for reductive dimethylation based quantitative phosphoproteomics (exemplified in a system of mouse liver phosphorylation) are summarized below. Our lab previously described, in detail, the work flow for a strong cation exchange-immobilized metal affinity chromatography (SCX-IMAC) strategy for large scale phosphopeptide enrichment[15]. In general this work flow is followed without deviation. The addition of an “on Sep-Pak” reductive dimethylation step prior to SCX was performed as described[16].

Starting material

Most phosphorylation events are of low stoichiometry, and many phosphoproteins are also of low abundance[17]. Thus, a large amount of starting material is required to obtain detectable levels of phosphopeptides. In addition, multidimensional enrichment steps (SCX and IMAC) are required to obtain a large data set by LC-MS/MS[15]. For these reasons, we began the analysis with 5 mg of peptides from fasted and re-fed mouse livers (a total of 10 mg after combining). This amount of material was sufficient for the identification and quantification of thousands of phosphorylation sites.

Mouse fasting and refeeding experiments

Mouse experiments were performed at the Dana Farber Cancer Institute with 8 week-old male Balb/c mice (Taconic). Mice were housed in sterile cages on a cycle of 12 hour light followed by 12 hour dark period and fed a standard chow diet (22.5% protein, 11.8% fat, 52% carbohydrate by mass). Mice were fasted overnight during the 12 hour dark cycle and sacrificed or fasted for 12 hours and re-fed beginning the next morning for 2 hours. Mice were sacrificed by CO_2 administration followed by cervical dislocation. Livers were removed by dissection and snap frozen using liquid nitrogen and stored at -80^{o}C until processing. Livers were ground using mortar and pestle, which was kept cold using liquid nitrogen both before and during pulverization prior to extraction of proteins.

Cell lysis

Lysates were prepared as previously described[15] in 8M urea buffered using 50 mM HEPES (pH 8.2) instead of Tris. This step was undertaken in order to avoid any possibility of the contaminating free amines of Tris from altering the dimethylation reaction efficiency. The buffer contained a cocktail of protease inhibitors (1 tablet Roche complete mini per 10 mL, 1 mM PMSF) and phosphatase inhibitors (1 mM NaF, 1 mM β-glycerophosphate, 1 mM sodium orthovanadate, 10 mM sodium pyrophosphate).

Cells were lysed and homogenized by sonication. Protein assays were performed using the BCA method. In total 4 lysates from fasted mice and 4 lysates from re-fed mice were prepared. 2 mg from each of the 4 fasted liver lysates were combined, and 2 mg from the 4 re-fed lysates were combined (thus reducing down to two separate samples, both with 8 mg of protein). Pooling lysates was used to control for stochastic biological variability. Yeast samples used for the pH titration experiment were prepared with the same buffer, lacking phosphatase inhibitors, and cells were lysed by bead beating.

Reduction, alkylation, and in solution digestion

In all cases, disulfide bonds were reduced in 5mM DTT at 56 °C for 45 min. Free sulfhydryl groups were then alkylated in 15mM iodoacetamide, in the dark at room temperature for 45 min. The reaction was quenched for 15 min at room temperature, in the dark with another addition of DTT, to the final concentration of 5 mM. The lysates were diluted 1:5 in 25 mM HEPES (pH 8.2), and calcium chloride ($CaCl_2$) as added to 0.1M to aid in digestion. Trypsin was added at 1:200 (enzyme: substrate ratio). Digestion reactions were incubated overnight at 37 °C. Digestions were acidified with TFA to a concentration of 0.5% (~pH 2) prior to solid phase extraction.

Peptide desalting and on-line reductive dimethylation of free amino groups

Peptides from the fasted and re-fed samples were separately loaded onto Sep-Pak cartridges (Waters Corporation) for desalting and reductive dimethylation labeling. Combining the desalting and labeling steps increase the throughput of the analysis. In both cases cartridges containing 200 mg of tC18 resin were used (so that the total amount of peptide loaded did not exceed 5% of the resin weight). The protocol for combining desalting and reductive dimethylation is as follows:

A vacuum manifold (Waters, #WAT200609) and a flow rate of 1-2 mL per minute were used for all steps except the binding and elution steps, which were undertaken without the use of an applied

vacuum. The flow rate was controlled in the binding and elutions steps by gravitational force alone. Prior to beginning this procedure, we prepared a solution of 0.4% formaldehyde (Sigma, St. Louis, MO) and 60 mM sodium cyanoborohydride (Sigma) in 0.25 M MES (pH 5.5), henceforth referred to as the "light reaction solution." Additionally, we prepared a solution of 0.4% deuterated (D2)-formaldehyde (Sigma) and 60 mM (D3)-sodium cyanoborodeuteride (CDN Isotopes, Canada) in 0.25 M MES (pH 5.5), henceforth referred to as the "heavy reaction solution." In our study, fasted liver peptides were labeled with the light reaction solution and are subsequently referred to as "light." Re-fed liver peptides were labeled with the heavy reaction solution and are subsequently referred to as "heavy." We followed the below procedure for each sample separately:

1. Wet the column with 6 ml of acetonitrile (ACN)
2. Wash with 6 ml 50% ACN/0.1% trifluoroacetic acid (TFA)
3. Equilibrate with 6 ml 0.1 % TFA
4. Load samples from 3.4 (peptides in 0.4% TFA)
5. Wash with 6 ml 0.1 % TFA
6. Wash column with 3 mL of 0.25 M MES (pH 5.5)
7. Label each sample by passing 10 mL of the either the heavy or light reaction solutions through the column, over the course of at least 10 min. Repeat this step to ensure complete labeling
8. The remaining steps should occur under a ventilated fume hood, as HCN is generated during quenching/washing.
9. Wash the column with 6 mL of 0.1% TFA
10. Wash the column with 1 mL of 0.5% acetic acid (AcOH) to remove TFA
11. Elute peptides with 2 mL of 70% ACN/0.5% AcOH
12. Combine the eluates and dry using vacuum centrifugation.

2-picoline-borane has been suggested as an alternative reductant to $NaBH_3CN$, which avoids HCN formation, though its use was not specifically addressed in this analysis. The use of pH 5.5 is a critical change from published protocols that avoids diminished identification rates seen at high pH. A summary of the reductive dimethylation chemistry, an example MS^1 spectrum of dimethylated peptides (light and heavy) and an extracted ion chromatogram of those peptides are presented below. The yeast samples used in pH comparisons were labeled in this manner, but with only the light form of

dimethylation reagents. The eluted yeast peptides from steps 11 were subsequently analyzed by LC-MS, producing the data which highlights this pH consideration.

Strong Cation Exchange (SCX) Chromatography

Details on the exact method for phosphopeptide separation by SCX have been previously described[15]. Briefly, 10 mg of the labeled heavy and light peptides were resuspended in 500 µL 7 mM KH_2PO_4, pH 2.65, 30% ACN (vol/vol) and separated into 12 equal (4 min) fractions on a polySULFOETHYL A column (9.4-mm inner diameter × 200 mm length, 5-µm particle size, 200 Å pore size, PolyLC). Peptides were loaded in 7 mM KH_2PO_4, pH 2.65, 30% ACN (vol/vol), buffer A, and eluted with 7 mM KH_2PO_4, 350 mM KCl, pH 2.65, 30% ACN (vol/vol), buffer B. A gradient of 100% A/0% B to 75% A/25% B was ran for 36 minutes, in order to effectively separate peptides. The flow rate was constant at 3 mL/minute. Each fraction was frozen in liquid nitrogen and lyophilized. The one experiment using yeast peptides (pH titration) did not involve the use SCX; further enrichment and cleanup steps were only required in the mouse liver phosphoproteome analysis.

SCX fraction desalting

Each SCX fraction was resuspended in 0.4% TFA (vol/vol). If the pH of a sample was greater than 2, an additional small volume of 10% TFA (vol/vol) was added until the sample pH was less than 2. The desalting occurred as described in 3.5, except for the use of 50 mg Sep-Paks instead of 200 mg Sep-Paks, all volumes were reduced by 50%, and no ReDi labeling step was carried out. Eluates were dried to completion using vacuum centrifugation.

Phosphopeptide enrichment using IMAC followed by combined elution and desalting on a STAGE tip

Prepare a STAGE tip[18] for each sample as follows:

1. Pack a 250 µL pipette tip with 2 disks of Empore 3M C18 material.
2. Wash each STAGE tip with 50 µL of 100 % methanol
3. Wash each STAGE tip with 50 µL of 50% ACN/1% FA, allow tip to remain wet for later
4. Set each STAGE tip off to the side and proceed with IMAC enrichment

IMAC resin (Phos-Select iron affinity gel; Sigma, St. Louis, MO) was equilibrated with three washes of 1% formic acid (FA)/40% ACN. Each desalted SCX fraction was resuspended in 100 µL of 1% FA/40% ACN and transferred to PCR tubes containing 20 µL of equilibrated IMAC slurry (1:1, beads:liquid, 10 µL beads). After a 60 min incubation (25 °C with vigorous shaking), the mixture was transferred to the prepared STAGE tips (above). Each STAGE tip was washed three times with 120 µL 1%FA/40% ACN to remove non-phosphorylated peptides from the IMAC resin (eluted peptides did not bind the STAGE tip resin). STAGE tips were equilibrated with a 50 µL wash of 1% FA. Phosphopeptides were eluted from IMAC beads with three 70 µL elutions of 500 mM dibasic sodium phosphate (K_2HPO_4), pH 7 (eluted peptides bound the STAGE tip resin). Phosphate salts were removed by washing the STAGE tips with 50 µL 1% FA. Phosphopeptides were eluted directly into glass inserts for analytical mass spectrometry using 40 µL 1% FA/50% ACN. The samples were dried to completion using vacuum centrifugation.

LC-MS/MS analysis and phosphopeptide identification

All LC-MS/MS analyses were performed on an LTQ Orbitrap XL hybrid mass spectrometer (Thermo Scientific, San Jose, CA). Dried phosphopeptide enriched samples were resuspended in 8 µL of 4% FA/5% ACN, and 4 µL were loaded onto a pulled fused silica microcapillary column (125 µm, 18 bed volume cm) packed with C_{18} reverse-phase resin (Magic C18AQ; 3-µm particles; 200-Å pore size; Michrom Bioresources, Auburn, CA) using a Famos autosampler (LC Packings, San Francisco, CA). Once loaded, the phosphopeptides were separated using an Agilent 1100 series binary pump (buffer A =

0.125% FA in 4% ACN, buffer B = 0.125% FA in 100 % ACN) across a 120 min linear gradient of 0% to 26% buffer B, at a flow rate of 500 nl/min. In each data collection cycle, one full MS scan (350-1800 *m/z*) was acquired in the Orbitrap ($3x10^4$ resolution setting at 400 *m/z*, automatic gain control (AGC) target of 10^6), followed by 10 data-dependent MS/MS scans in the LTQ (AGC target of 5,000; minimum threshold of 3,000) using the 10 most abundant ions and collision-induced dissociation (CID) for fragmentation[19].

Once selected, ions were dynamically excluded for 60s. Singly charged ions and unassigned charge states were always excluded. Maximum ion accumulation times were 1000 ms for each full MS scan and 150 ms for MS/MS scans. Lockmass, using the atmospheric contaminant polydimethylsiloxane (m/z 371.1012) as an internal standard was used to correct precursor ion *m/z* values. Samples were shot twice on the mass spectrometer to increase proteome coverage and assist in assessing quantitative reproducibility. Yeast samples discussed in the pH titration experiment were collected in a similar manner, however a 65 min gradient was used in place of a 120 min gradient, and selected ions were dynamically excluded for 30s instead of 60s. The longer gradient was only required to maximize phosphopeptide identifications.

RAW files obtained from data collection were converted into mzXML format using the ReAdW program (http://sourceforge.net/projects/sashimi/files/ReAdW%20%28Xcalibur%20converter%29/). The SEQUEST search algorithm (version 28 [20]) was used to search MS/MS spectra against a composite database comprised of all mouse (or yeast where relevant) open reading frames in their forward and reversed orientations [21]. The search parameters used are as follows: 25 ppm precursor ion tolerance and 1.0 Da fragment ion tolerance; fully tryptic digestion; up to two missed cleavages were allowed; posttranslational static modifications of 57.02146 Da on cysteine (carboxyamidomethylation) and 28.03130 Da on lysine and the peptide N-terminus (dimethylation, light); dynamic modifications of 15.99491 Da on methionine (oxidation), 79.96633 Da on serine, threonine, and tyrosine (phosphorylation, only for mouse searches) and 6.03766 on lysine and the peptide N-terminus

(dimethylation heavy, only for mouse searches). Yeast peptides from the pH titration experiment were searched in a similar manner, though only a dynamic modification methionine oxidation was included, as only the light form of the dimethylation reagents were used, and the samples were not enriched for phosphopeptides. The labeling efficiency of these yeast samples was assessed by including dynamic modifications of 14.0157 and 28.03130 Da on lysine and the N-terminus in the search criteria, while carboxyamidomethylation (57.02146 Da on cysteine) was the only static modification allowed. Using such criteria for the search, unlabeled, partially labeled and fully labeled peptides could be identified. The dynamic modification of oxidation (15.99491 Da on methionine) was still allowed as well in these searches.

Data filtering and phosphorylation site localization.

In all cases, matched peptide spectra were first filtered using a target-decoy strategy [21] to a 1% peptide level false discovery rate (FDR) through linear discriminant analysis (LDA) using the following parameters: XCorr, ΔCn', precursor mass error, solution charge (when analyzing SCX fractions), charge state, number of missed cleavages, and regular expressions for complete labeling and phosphorylation (when applicable) [1]. ΔCn' is defined by the XCorr difference between the top SEQUEST hit and that of the first subsequent unique sequence hit (not simply phosphorylation site placement), divided by the XCorr of the top hit. Linear discriminant models were calculated for each run using peptide matches to forward and reversed protein sequences as training data. Peptides in each MS/MS run were ranked by descending discriminant score and filtered to a 1% FDR based on the number of reverse sequences in the data set. The data was subsequently filtered to control the protein level FDR. Protein scores were created from LDA peptide probabilities, sorted by rank, and filtered to 1% FDR as described for peptides above[1]. Any time a group of samples is simultaneously considered, new filters must be created to properly asses the protein level FDR. For example unique protein level filters were created for each set

of 12 technical replicates alone (two groups of 12 samples), and for the combination of all data at once (one group of 24 samples). The Ascore algorithm was used to assign phosphorylation site localizations, with a score of 13 ($p<0.05$) considered to be localized[22].

Phosphorylation site quantification

Heavy to light dimethylated peptide ratios were generated for each peptide by extracting ion chromatograms (XICs) within 10 ppm of the observed m/z for the monoisotopic and first ^{13}C peaks for both heavy and light isotopes, as previously described[23]. Each chromatogram was integrated and the heavy XIC was divided by the light XIC to determine the dimethylation ratio. Signal-to-noise (S/N) ratios were determined by comparing their observed signal intensities with the median signal intensity observed in nearby m/z ranges encompassing several minutes around a peptide's elution. In cases where only a single isotopic species (heavy or light) was present, the S/N ratio was used for quantitation instead. Peptides were required to have either a minimum heavy and light S/N of 5, or a S/N of 10 for one of the two isotopic species (with no requirement for the other isotopic species), in order to be considered for quantification. An example of quantified heavy and light XICs is shown in below. All ratios discussed are relative to the heavy sample, so that upregulated site are more abundant in the re-fed livers and downregulated sites are more abundant in the fasted livers.

Motif analysis

General motif classes/sequence categories were assigned given the rules previously defined[1]. Specific motifs were extracted from the data set using the Motif-X algorithm[24] (http://motif-x.med.harvard.edu). Sequences were centered at the phosphorylated residue and extended 6 amino acids on each side, giving a total length of 13 amino acids for each phosphorylation site. Sites with an Ascore >13 were used for motif extraction, not including N/C-terminal peptides. The minimum reported

number of occurrences for a given motif was set at 20. Only motifs with a motif score of >6 (binomial probability $<10^{-6}$) were reported. Sequence logos were automatically generated by the Weblogo program[25] (http://weblogo.berkeley.edu).

Results and Discussion

Stable Isotope Incorporation by Reductive Dimethylation Allows for the Identification and Quantification of Peptides by Mass Spectrometry

An example of the reductive dimethylation reaction is highlighted in Figure 3.1. In the reaction, free amino groups (lysine and N-terminus) are dimethylated with formaldehyde using a sodium cyanoborohydride reductant. Incorporation of stable deuterium isotopes in both reagents allows for heavy labeling of peptides. There is a +6 Da shift between heavy and light peptides, per free amino group, using the outlined reagents. Based on the chemistry, two of the three hydrogen/deuterium atoms in each methyl group (as well as the carbon atom) originate from the formaldehyde/D2-formaldehhyde and one originates from the NaBH3CN/NaBD3CN (Figure 3.1A). Thus different combinations of reagents (including the use of ^{13}C D2-formaldehyde) allow for three-plex[12] and even five-plex reductive dimethylation (data not shown); however, increasing the complexity of the sample by introducing additional isotopic forms of a peptide hinders the comprehensiveness of the analysis (data not shown). The precursor ion spectrum for the peptide FENAFLSHVISQHQSLLGNIR (3+) is presented in Figure 3.1B. There was a 2 *m/z* shift between the peptides (6 Da/3 charges) and the isotopic envelopes for both species were amply resolved. Extracted ion chromatograms from the ions in part B were used to calculate a heavy to light ratio for the peptide (Figure 3.1C). In the example, the difference in the ratio obtained by peak height vs. peak area integration was ~15%. In some cases, with more pronounced

differences in the elution profiles of the heavy and light peptides (due to deuterium effects), errors in ratio calculation between peak height and integration would be much greater. All reported values use area integration to avoid errors in quantification.

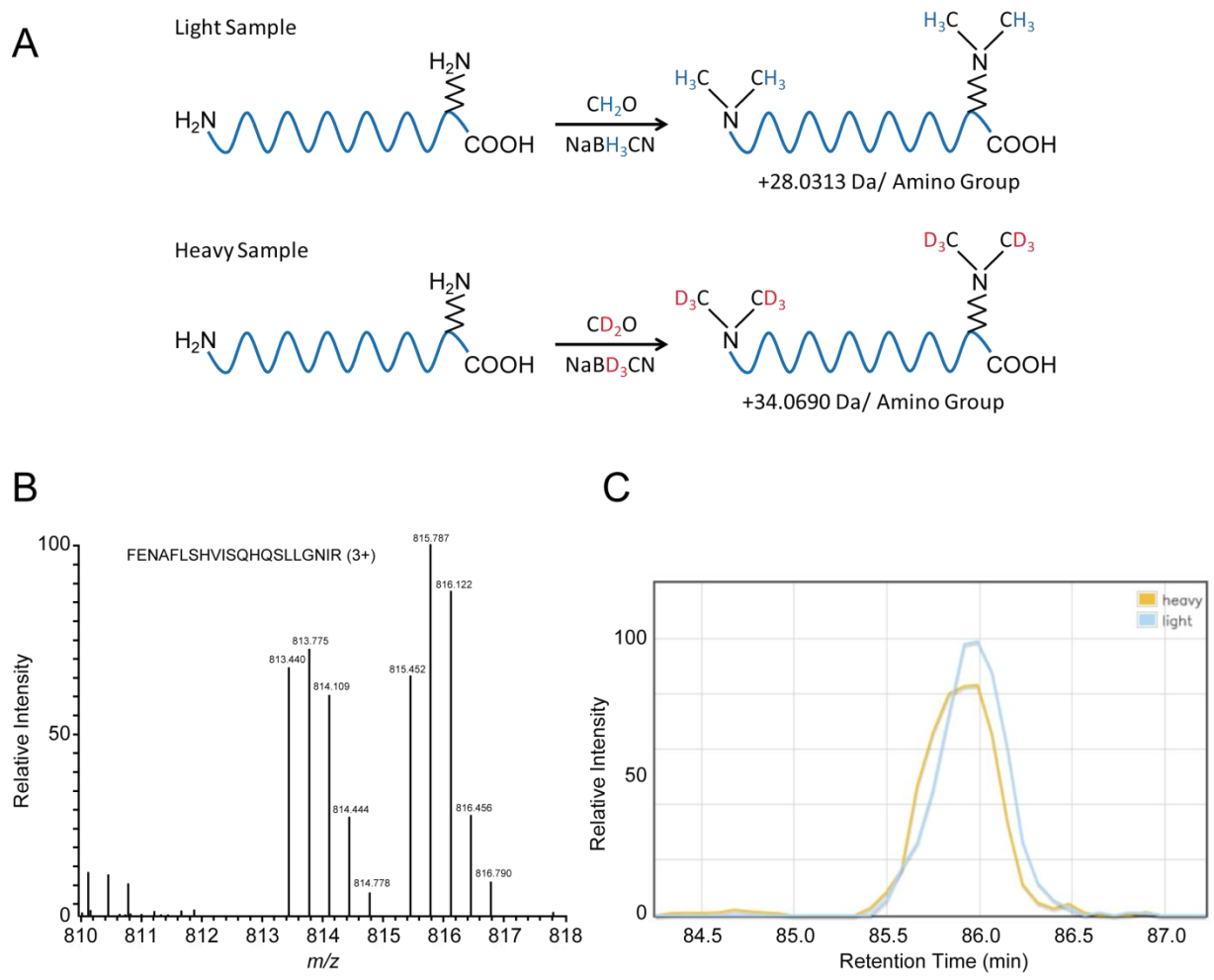


Figure 3.1. The physical characteristics of reductively dimethylated peptides are amenable to quantitative mass spectrometry. Representation of the peptide reductive dimethylation reaction using heavy and light reagents (A). Reductive dimethylation occurs on free amino groups, labeling the N-terminus and any lysine side chains of a peptide. Paired peptides using the given reagents display a 6.0377 Da mass shift in the mass spectrometer with respect to one another (per label, divided by charge). Example MS^1 ions for the peptide FENAFLSHVISQHQSLLGNIR (3+), labeled on the N-terminus only (B). The Orbitrap provided ample resolution in order to distinguish isotopic envelopes for each peptide. The extracted ion chromatograms for both the heavy and light peptides from the ions in part B are shown (C). Though there is a deuterium effect on the elution profiles (slight retention time shift), by integrating the area of the ion chromatograms and not relying on maximal peak intensity alone, one can accurately assess the heavy to light ratio.

Lower pH Conditions Are Required for Successful Peptide Reductive Dimethylation

A wide pH range is given (5-8.5) as the optimal pH for the reaction, and commonly the reaction is performed at pH 7.5 or 8 in the literature [12, 26]. We found, however, a tighter pH range (5-6) to be optimal for peptide identification. To evaluate pH as a factor in peptide identification rates, we dimethylated trypticly digested yeast whole cell lysate on a Sep-Pak column at different pHs, prior to LC-MS/MS (Figure 3.2A). Lysis, digestion and labeling methods are as described above. A pH range of 4-8 was chosen to demonstrate the trend between dimethylated peptide identification rates and reaction pH. Only light reaction solutions (prepared at the proper pH) were used for this analysis. We found that with increasing reaction pH, the identification of total peptides, unique peptides and proteins was significantly reduced. The maximum identifications were observed at pH 5 (4033 total peptides, 3150 unique peptides, and 808 proteins) and minimum identifications were observed at pH 8 (1760 total peptides, 1360 unique peptides, and 442 proteins, Figure 3.2C). Both the base peak elution chromatograms and the total number of MS/MS collected did not change significantly between the different reaction conditions, suggesting that both the LC and MS components of the analysis were equal in all cases (Supplemental Figure 3.1 and Figure 3.2C, respectively). XCorr values for matched peptides were also comparable between reaction conditions (Supplemental Figure 3.2), suggesting differences in identifications were not due to fragmentation efficiency.

Potential causes for the ~60% reduction in identifications are altered labeling efficiency and/or side product formation. Indeed, at pH8 for example, labeling efficiency was only ~85% based on searches performed with dynamic modifications for partial and full labeling (+14.0157 and +28.03130 Da on the peptide N-terminus and lysine residues, Supplemental Figure 3.3). Though this result explains some of the noted difference, it alone is not sufficient and suggests additional reaction products. In fact at high pH, many peptides were found to be modified by 24.995 AMU, matching a cyano group (in place of hydrogen) addition (based on mass from www.unimod.org, Supplemental Figure 3.4). As $NaBH_3CN$

was used as a reagent, this mass addition is possible. These data have been replicated on several occasions, and the pH dependence of the reaction holds true whether it occurred on column or in solution (data not shown). This pH consideration was an important finding for the application of reductive dimethylation to a proteomic scale.

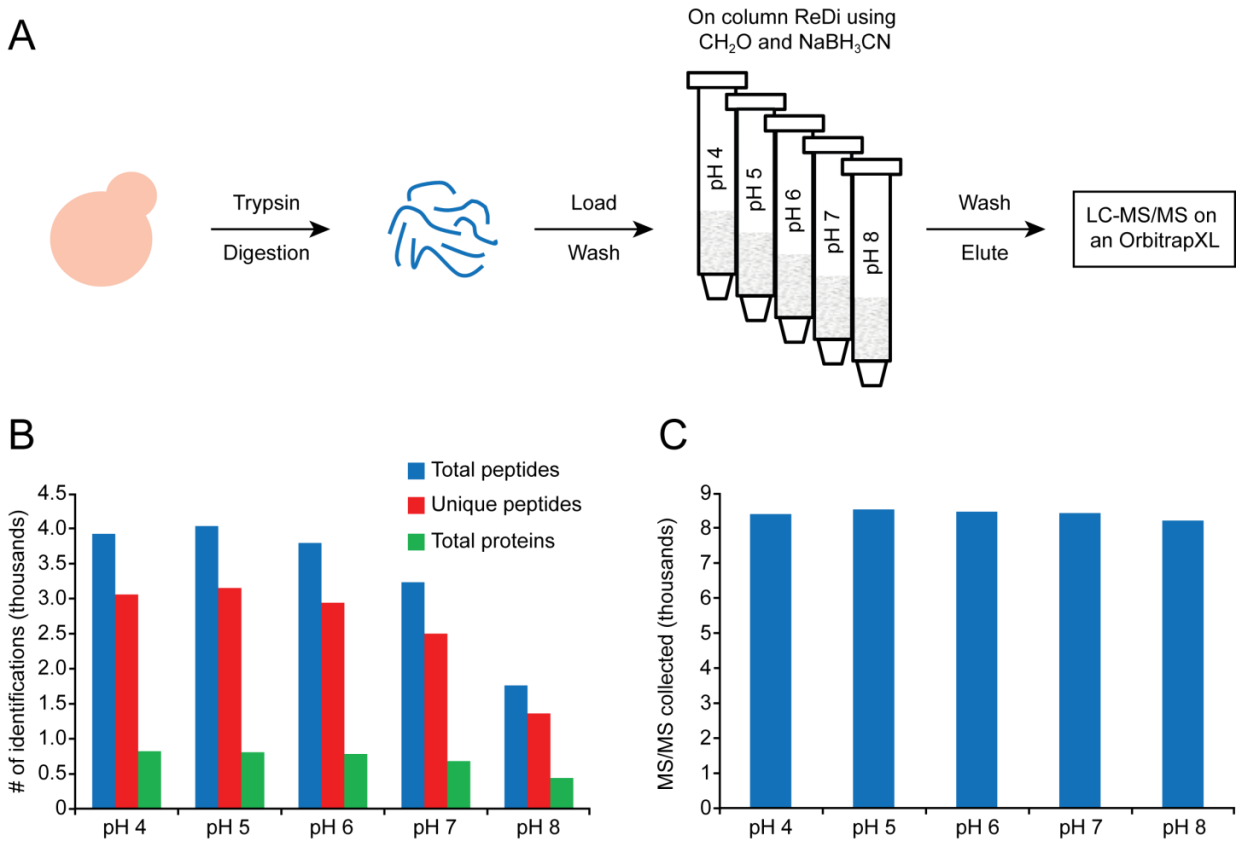


Figure 3.2. Higher reductive dimethylation reaction pH yields fewer dimethylated peptide and protein identifications. Yeast whole cell lysate was prepared and digested using trypsin (as described in the methods section). Aliquots of unlabeled peptides were separately loaded on to five tC18 Sep-Pak columns (A). The reductive dimethylation reactions, using only light reagents, were carried out at pH 4-8 to assess the relationship between pH and identification success rates. All data were analyzed on an LTQ Orbitrap XL by LC-MS/MS over a 65 minutes gradient. A clear trend between increasing pH and lower success of peptide/protein identification was detected (B). These data show that a pH between 5 and 6 is optimal for the dimethylation of peptides. Similar numbers of total MS/MS spectra were collected in all cases (C).

Thousands of Phosphopeptides Are Identified and Quantified Using a Lower pH Reductive Dimethylation Strategy in Fasted vs. Re-Fed Mouse Liver Samples

Using a lower pH reductive dimethylation strategy as outlined in Figure 3.3, thousands of phosphopeptides were successfully identified and quantified in fasted and re-fed mouse liver samples. Data were collected as technical replicates (replicate LC-MS analyses). Dataset statistics for both replicates (12 SCX fractions) and the combined data set (2 replicates of all 12 fractions) filtered to a 1 % peptide and protein level FDR are presented in Table 3.1 (bottom). The same data without protein level filtering is also presented in Table 3.1 (top) to highlight the need to control protein level FDR when analyzing quantitative data (discussed below). In both tables the FDR at each level is listed in parenthesis. Only the data set using peptide and protein level filters was used for further analyses. In total ~38, 000 redundant phosphopeptides from 2741 phosphoproteins were identified. Nearly 12,000 unique phosphopeptides (sequences stripped of all modification except phosphorylation) and >8500 unique phosphorylation sites were identified. Nearly 7400 sites from >2300 phosphoproteins were reliably quantified (passing the describe filter criteria, see materials and methods). To facilitate the analysis, sites that changed by 2-fold were considered to be regulated including 210 downregulated and 180 upregulated sites. These data comprise one of the largest phosphorylation analyses using reductive dimethylation and were acquired in less than 4 days of instrument time. All identified peptides and sites are presented in supplemental Table 3.1.

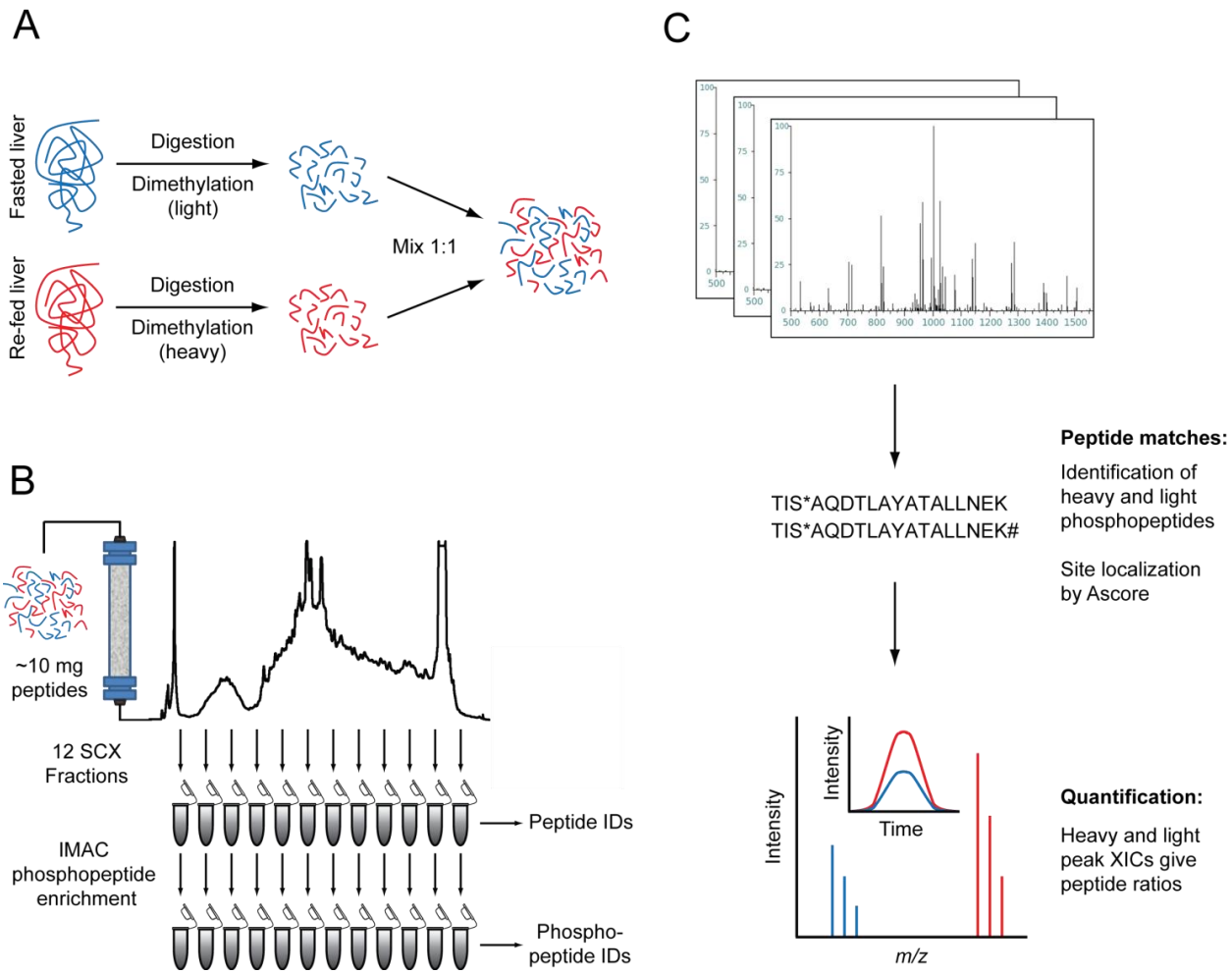


Figure 3.3. Lower pH reductive dimethylation allows for the successful quantification of thousands of phosphorylation sites, using an SCX/IMAC strategy[15]. Pooled lysates from four fasted (harvested after a 12 hour fast) and four re-fed (12 hour fast followed by a two hour re-feed prior to harvesting) were reduced, alkylated and digested by trypsin (A). Peptides from the fasted liver pool were labeled light and peptides from the re-fed liver pool were labeled heavy using reductive dimethylation. Mixed (1:1) peptides were separated by strong cation exchange prior to IMAC enrichment for phosphopeptides (B). Phosphopeptide-enriched samples were analyzed by LC-MS/MS on an LTQ-Orbitrap XL mass spectrometer. MS/MS spectra were searched and assigned to peptides by SEQUEST[20] (C). Phosphopeptides were filtered to a desired peptide and protein level false discovery rate (FDR <1%) and site localization was assessed using the Ascore algorithm[1, 22]. Extracted ion chromatograms corresponding to the heavy and light peptides from which each identified site were derived were used to obtain quantitative site ratio between the fasted and re-fed mice.

Protein Level Filters Are Required to Control Both False Discovery Rates at the Protein Level and Phosphorylation Site Level

In many published quantitative phosphorylation datasets, false discovery rates are controlled at the peptide level only. As is highlighted in Table 3.1 (top), combining multiple LC-MS runs, even with

peptide level filtering, greatly increased both protein and site level false discovery rates. Remarkably, the peptide level FDR could still be reported at 1%, yet the protein level FDR in each set of technical replicates expanded to 6%. When all MS/MS runs were combined, the protein level FDR became >10%. Site level FDRs showed a similar trend, though due to Ascore filtering, the FDR was slightly reduced compared to that of the protein level FDR. With protein level filters in place, a desired FDR of 1% could be set within each set of replicates or in all runs combined (Table 3.1 bottom). As a result the FDR of identified sites was <0.7%, roughly 9 fold less than without protein level filters. Peptide identifications as a whole were minimally affected by protein level filtering; generally the excluded proteins were those which were identified based upon single peptide observations.

What is of particular concern, however, is that the false positive hits tended to cluster in sites that changed by ≥2 fold (regulated sites). One reason for this phenomenon is an inherent limitation of MS^1-based quantification; namely that the quantification of a peptide is tied directly to the identification of a peptide. A false positive hit (whether a forward sequence or reverse sequence) has an equal chance of being assigned heavy or light. If a peptide spectral match that in reality is a light labeled peptide is matched instead to a sequence that is heavy labeled, the quantification of that peptide will be erroneous, as it is likely that either no peak or an unrelated peak will be chosen as its isotopomer. This error results in a site quantification value that is often greater than two fold (either upregulated or downregulated). Without any protein level filtering, the FDR of regulated sites (all 24 MS/MS runs combined) reached 40%. Our FDR estimates for site quantification are approximations and only account for misidentifications; these estimates may not fully account for mislocalized sites on otherwise correctly identified peptides. They are, however, a useful guide to understanding erroneous identification/quantification on the site level.

Table 3.1. Fasted vs. re-fed mouse liver data set statistics, after controlling only peptide level (top) or both peptide level and protein level false discovery rates (bottom). Replicates refer to replicate LC-MS analyses of each SCX fraction. Data are shown by both individual technical replicates on the mass spectrometer (12 MS runs) and in a combined data set (24 MS runs). Estimated false discovery rates are given in parentheses based on the target-decoy approach[21]. 558 sites were found to change by at least two-fold, at a false discovery rate of 40% when using only peptide filters. 390 sites were found to change by at least two-fold, at a false discovery rate of <6% (at the regulated site level, <0.5% for all quantified sites) when using both peptide and protein filters. Site level FDRs are approximations.

Peptide level filters only

	Replicate one	Replicate two	Combined
MS/MS	130 304	123 348	253 652
Total phosphopeptides	19 967 (1%)	19 006 (1%)	38 973 (1%)
Phosphoproteins	2 863 (6%)	2 891 (6%)	3 296 (10%)
Phosphorylation sites	7 411 (4%)	7 553 (4%)	9 239 (6%)
Downregulated sites (2 fold)	227 (38%)	213 (31%)	310 (46%)
Upregulated sites (2 fold)	176 (25%)	208 (21%)	248 (33%)

Peptide and protein level filters

	Replicate one	Replicate two	Combined
MS/MS	130 304	123 348	253 652
Total phosphopeptides	19 592 (0.2%)	18 609 (0.2%)	38 314 (0.1%)
Phosphoproteins	2 530 (1%)[a]	2 504 (1%)[a]	2 741 (1%)[b]
Phosphorylation sites	6 990 (0.6%)	7 078 (0.6%)	8 540 (0.7%)
Downregulated sites (2 fold)	167 (10%)	162 (5%)	210 (6%)
Upregulated sites (2 fold)	144 (3%)	165 (2%)	180 (4%)

[a] Protein level FDR controlled in each replicate data set (12 SCX fractions)
[b] Protein level FDR controlled in all runs combined (24 SCX fractions)

The protein level FDR was set to 1% at either the technical replicate level or for the entire combined dataset. The protein level FDR must be controlled within each group of samples being considered to accurately estimate the FDR. This filtering lowered the FDR for regulated sites to <6%, and

<0.5 % for all quantified sites. A separate study aimed at controlling the FDR of regulated sites in a quantitative phosphoproteomic study would be of great value to the proteomics community. Additional filters which helped to reduce the regulated site level FDR are presented in Supplemental Figure 3.5. Though filters based on the number of peptides used to quantify a site and filters based on increasing Ascore were valuable in the reduction of the FDR, they did so at a substantial cost to site identification rates. To a lesser extent, filters using S/N and Z-score (number of standard deviations from the mean) of the heavy to light ratio had an impact on identifications while controlling the regulated site level FDR. Rather than use any of the aforementioned values as a hard cutoff, they may be used as a guide for the confidence in quantification of a given site. For example if a site was quantified using several peptides, with S/N for both heavy and light species greater than 3, it was likely a correct quantification.

Reductively Dimethylated Peptides Follow Known Trends for Large-Scale Phosphoproteomic Analyses

With any labeling method, there is concern that the label will affect the quality of data. The dataset presented in this study is of sufficient size, thus enabling it to be compared with other large-scale phosphoproteomics data sets (for example[7] and [1]). Figure 3.4A highlights the identification of phosphopeptides, phosphorylation sites and phosphoproteins amongst SCX fractions for one set of technical replicates. Dimethylation does not affect the elution of phosphopeptides from the SCX column, as the data are consistent with other published studies [7, 15]. Earlier SCX fractions were enriched for multiply phosphorylated peptides (Figure 3.4B). The first two fractions for example contained >90% multiply phosphorylated peptides. This observation is of particular importance as these peptides significantly increase the number of sites that are reported in large scale phosphoproteomic datasets. Technical replicates tend to increase the number of non-redundant site identifications by 15-20% (Figure 3.4C).

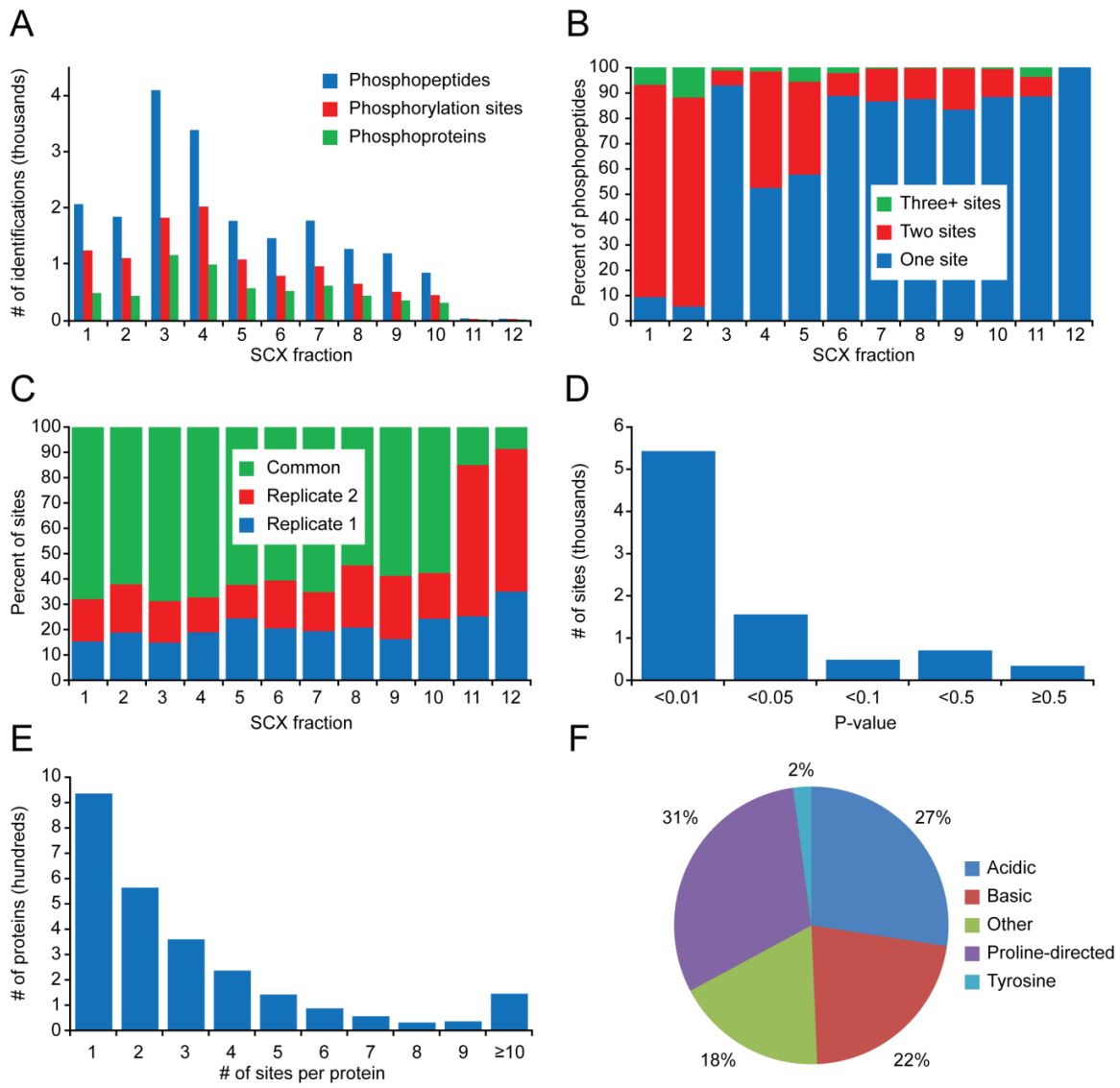


Figure 3.4. Phosphopeptide and phosphorylation site data are consistent with known trends for large scale phosphorylation datasets. Data obtained using lower pH reductive dimethylation showed similar trends to both unlabeled [7, 15] and SILAC-labeled[10] datasets. The majority of phosphopeptides eluted in earlier SCX fractions (A) and the number of phosphorylation sites per peptides was negatively correlated with SCX fraction number (B). By using technical replicates, the number of identified phosphorylation sites was increased by ~15-20% per fraction (C). The vast majority (~82%) of sites were localized with high certainty ($p < 0.05$, D). As observed in other studies [1], most identified phosphoproteins were multiply phosphorylated (~65 %, E). Proline-directed phosphorylation composed the largest fraction of phosphorylation site data (31%), followed by acidic (27%) and basic (22%) general motif classes (F). General motif classes have been previously defined in detail[1].

The vast majority of non-redundant identified sites were localized with high certainty (82%, $p<0.05$, 3.4D). This observation suggests that the dimethylation of phosphorylated peptides does not affect their fragmentation efficiency in the ion trap, or negatively affect neutral loss of phosphate; thus,

many site determining ions are produced. The majority (~65%) of phosphoproteins identified in this study were multiply phosphorylated, and a large number (>100) were phosphorylated on ≥10 residues (Figure 3.4E). General motif class distributions including acidic (27%), basic (22%), proline-directed (31%), uncharacterized/other (20%) and tyrosine (2%) were consistent with other large scale studies (Figure 3.4F)[1, 7]. The majority of identified sites (84%) were on serine residues, followed by threonine (14%) and tyrosine (2%). These data taken together show that reductive dimethylation coupled to SCX/IMAC provides a robust strategy for large scale phosphoproteomics.

Analysis of Quantitative Site Data

Quantitative site ratios were $\log_2$ transformed and adjusted for mixing errors by normalizing the site ratios so that the median ratio equaled 0. This procedure effective re-centered the distribution of site ratios (Figure 3.5A). The majority of sites quantified in this study fell close to a 1:1 (heavy to light) ratio, 390 sites changed by a two-fold (one $\log_2$ unit) or greater ratio. These sites were considered to be regulated. A site whose $\log_2$ ratio was significantly greater than 0 was more abundant upon re-feeding mice, where a site quantified with a ratio significantly less than 0 was more abundant in the fasted mice. A histogram of site ratios is presented in Figure 3.5A, with arrows indicating two-fold changes. The majority of phosphorylation sites were quantified using at least two peptides for quantification (Figure 3.5B). In those cases, the reported value for a site is the median heavy to light ratio for all peptides from which the site was derived. A plot of the summed signal-to-noise ($\log_2$) vs. heavy to light ratio ($\log_2$) for each site (Figure 3.5C) shows that many of the regulated sites (at least two fold change, blue triangles) were quantified with good S/N measurements (>10), an indicator for quality of quantification; sites which changed by less than two fold are plotted as small black circles. Many of the sites which changed by larger values (>4 fold) also are quantified at high summed S/N. On the whole, site quantification between technical replicates was reproducible (slope = 0.95, R^2=0.78, Figure 3.5D). When sites that

changed by less than two fold were removed (Supplemental Figure 3.6), the slope of the linear regression line and the coefficient of determination approached 1. It is likely that small deviations in the large number of data points quantified at close to a 1:1 (heavy : light) ratio adversely affected the correlation in Figure 3.5D.

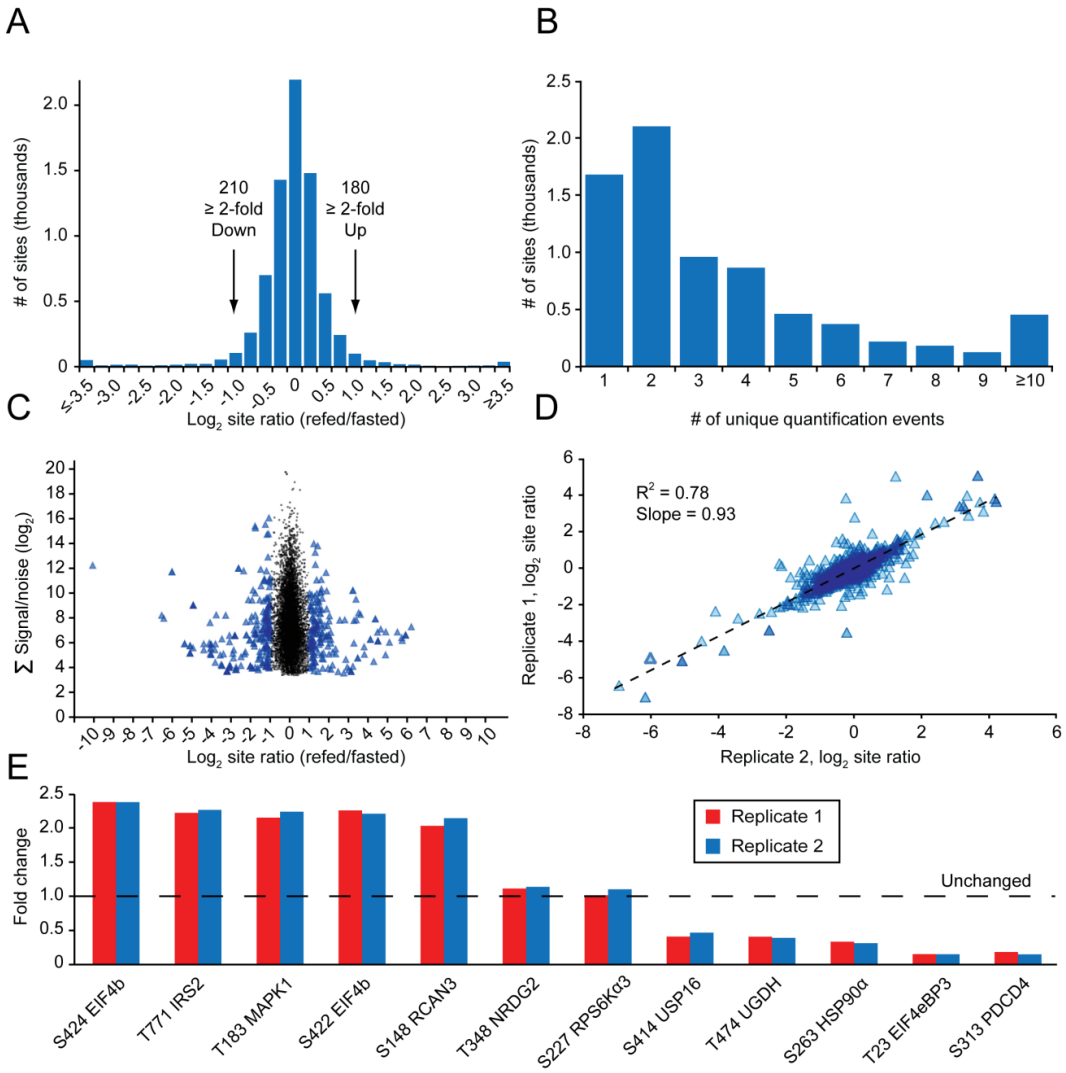


Figure 3.5. Quantitative phosphorylation site data. The heavy to light phosphorylation site ratios (log_2) are plotted as a histogram (A). Sites showing a 2-fold change were considered to be regulated. The majority of phosphorylation sites (~80%) were quantified by multiple unique quantification events (B). Site ratios are plotted against the summed signal-to-noise for all peptides encompassing that site (C). All peptides used to quantify site were required to have a minimum S/N of 5 for both the heavy and light peptides or a S/N of 10 for one of the peptides (heavy or light). Site ratios were reproducible on a whole between technical replicates, based on linear regression of the observed site ratio between replicates (D). Highlighted phosphorylation site ratios (unlogged) for example proteins (protein residue locations are indicated) observed in each replicate are plotted (E), see text for details. The dashed line indicates a 1:1 ratio (heavy : light). Replicates signify replicate LC-MS runs, and represent analytical stability.

Reductive Dimethylation Provides Biologically Relevant Quantitative Phosphorylation Data

Figure 3.5E highlights example phosphorylation sites from the data set. The median unlogged heavy/light ratio for each site is plotted by technical replicate, with the protein abbreviation and phosphorylation residue number indicated under each column. We discovered an upregulated site on the eukaryotic translation initiation factor EIF4b (S422, located in the RNA binding region), which had previously been proposed to be phosphorylated by the 90 kDa ribosomal protein S6 kinase 1 (RSK1), and required for translation [27]. Additional sites on EIF4b including the nearby S422 were also upregulated in our dataset. An activating phosphorylation site on RSK2 (PDK1 mediated[28], S227 RPS6K3α), however, was unchanged here. These data suggest that RSK1 and not RSK2 may be activated after 2 hours of refeeding, and is perhaps responsible for the associated changes in signal transduction. Such observations warrant confirmation of the observed behavior and further analysis into the temporal dynamics of RSK isoform-specific signaling. Sites on other important metabolic proteins such as insulin receptor substrate 2 (T771 IRS2) and map kinase (T183 MAPK1, a canonical activation site[29]) were also found to be upregulated.

Interestingly, a phosphorylation site in the FLISPP motif of calcepressin-3 (S148 RCAN 3) was found to be upregulated. The calcipressin family of proteins is known to inhibit calcineurin mediated signaling, a calcium/calmodulin-dependent phosphatase [30]. Phosphorylation of this domain in the related protein calcepressin-1 has been shown to increase its ability to inhibit calcineurin [31]. Furthermore it was recently reported that in fasted mice, glucagon promotes the release of intracellular calcium stores, thereby activating calcineurin[32]. Calcineurin is then able to dephosphorylate and activate CRTC2 (A CREB coactivator) and execute gluconeogenic programing. This same study also showed that insulin signaling leads to the deactivation of CRTC2. Our observation that re-feeding leads to the increase of a phosphorylation site that may play a role in regulating calcineurin activity is consistent with this model. Finally a phosphorylation site on NMYC downstream-regulated gene 2 (T348 NRDG2), a

protein linked to cell proliferation, had been previously shown to be phosphorylated by Akt [33], but showed no change in our analysis. Potentially both canonical and non-canonical signaling cascades are occurring in these mice.

The highlighted downregulated sites in Figure 3.5E were less clear as to their potential function, but are nonetheless interesting based on the function of the proteins on which they were found. USP16 is a deubiquitinase involved in H2A deubiquitination and is required for cell division [34]. UGDH is responsible for the synthesis of glycosaminoglycans and thus the maintenance of the extracellular matrix. HSP90α (a cochaperone) is responsible for the proper maintenance of proteins involved in the cell cycle and signal transduction, for example. The protein EIF4eBP3 may be directly involved in the regulation of translation through its inhibition of the EIF4F complex (by binding the EIF4e subunit [35]). Finally Programmed cell death protein 4 (PDCD4) is also involved the inhibition of translation initiation and may be involved in apoptosis [36].

An analysis of general and sequence specific motifs between the whole data set and the down and upregulated sites is presented in Supplemental Figure 3.7. We identified changes in several known motifs including basic, proline-directed, and acid motifs (see Figure legend for details).

Conclusions

In this publication, we present a recommendation for a lower reaction pH as an improvement to the reductive dimethylation protocols. We applied the amended protocol to the large-scale quantitative analysis of phosphorylation sites from mouse liver tissue, comparing fasted and re-fed states. This model system provided a framework to demonstrate the use of reductive dimethylation for identifying potentially relevant phosphorylation events; in this example, those sites potentially involved in energy homeostasis were highlights. In addition, the size of the generated data set was sufficient to remark on

the effects that peptide and protein level FDR filtering have on MS^1 based-quantification. It was demonstrated that permissive filtering greatly increases the FDR of the subset of regulated phosphorylation sites (those which changed by 2-fold or greater). This publication provides a framework for future quantitative phosphoproteomic endeavors, particularly for the analysis of mammalian tissues.

Acknowledgments

This work was supported in parts by grants awarded to S. P. G from the National Institutes of Health (NIH; HG3456 and GM67945). We would like to Dr. Timothy Kelly and Dr. Pere Puigserver for providing the fasted and re-fed mouse liver samples.

References

1. Huttlin, E. L.; Jedrychowski, M. P.; Elias, J. E.; Goswami, T.; Rad, R.; Beausoleil, S. A.; Villen, J.; Haas, W.; Sowa, M. E.; Gygi, S. P., A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **2010,** 143, (7), 1174-89.
2. Scanff, P.; Yvon, M.; Pelissier, J. P., Immobilized Fe3+ affinity chromatographic isolation of phosphopeptides. *J Chromatogr* **1991,** 539, (2), 425-32.
3. Beausoleil, S. A.; Jedrychowski, M.; Schwartz, D.; Elias, J. E.; Villen, J.; Li, J.; Cohn, M. A.; Cantley, L. C.; Gygi, S. P., Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A* **2004,** 101, (33), 12130-5.
4. Larsen, M. R.; Thingholm, T. E.; Jensen, O. N.; Roepstorff, P.; Jorgensen, T. J., Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol Cell Proteomics* **2005,** 4, (7), 873-86.
5. Wilson-Grady, J. T.; Villen, J.; Gygi, S. P., Phosphoproteome analysis of fission yeast. *J Proteome Res* **2008,** 7, (3), 1088-97.
6. Zhai, B.; Villen, J.; Beausoleil, S. A.; Mintseris, J.; Gygi, S. P., Phosphoproteome analysis of Drosophila melanogaster embryos. *J Proteome Res* **2008,** 7, (4), 1675-82.
7. Villen, J.; Beausoleil, S. A.; Gerber, S. A.; Gygi, S. P., Large-scale phosphorylation analysis of mouse liver. *Proc Natl Acad Sci U S A* **2007,** 104, (5), 1488-93.
8. Kruger, M.; Moser, M.; Ussar, S.; Thievessen, I.; Luber, C. A.; Forner, F.; Schmidt, S.; Zanivan, S.; Fassler, R.; Mann, M., SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. *Cell* **2008,** 134, (2), 353-64.
9. Nagaraj, N.; D'Souza, R. C.; Cox, J.; Olsen, J. V.; Mann, M., Feasibility of large-scale phosphoproteomics with higher energy collisional dissociation fragmentation. *J Proteome Res* **2010,** 9, (12), 6786-94.

10. Dephoure, N.; Zhou, C.; Villen, J.; Beausoleil, S. A.; Bakalarski, C. E.; Elledge, S. J.; Gygi, S. P., A quantitative atlas of mitotic phosphorylation. *Proc Natl Acad Sci U S A* **2008,** 105, (31), 10762-7.
11. Gruhler, A.; Olsen, J. V.; Mohammed, S.; Mortensen, P.; Faergeman, N. J.; Mann, M.; Jensen, O. N., Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol Cell Proteomics* **2005,** 4, (3), 310-27.
12. Boersema, P. J.; Raijmakers, R.; Lemeer, S.; Mohammed, S.; Heck, A. J., Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat Protoc* **2009,** 4, (4), 484-94.
13. Zoumaro-Djayoon, A. D.; Ding, V.; Foong, L. Y.; Choo, A.; Heck, A. J.; Munoz, J., Investigating the role of FGF-2 in stem cell maintenance by global phosphoproteomics profiling. *Proteomics* **2011,** 11, (20), 3962-71.
14. Lemeer, S.; Jopling, C.; Gouw, J.; Mohammed, S.; Heck, A. J.; Slijper, M.; den Hertog, J., Comparative phosphoproteomics of zebrafish Fyn/Yes morpholino knockdown embryos. *Mol Cell Proteomics* **2008,** 7, (11), 2176-87.
15. Villen, J.; Gygi, S. P., The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat Protoc* **2008,** 3, (10), 1630-8.
16. Tolonen, A. C.; Haas, W.; Chilaka, A. C.; Aach, J.; Gygi, S. P.; Church, G. M., Proteome-wide systems analysis of a cellulosic biofuel-producing microbe. *Mol Syst Biol* **2011,** 7, 461.
17. Wu, R.; Haas, W.; Dephoure, N.; Huttlin, E. L.; Zhai, B.; Sowa, M. E.; Gygi, S. P., A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nat Methods* **2011,** 8, (8), 677-83.
18. Rappsilber, J.; Mann, M.; Ishihama, Y., Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* **2007,** 2, (8), 1896-906.
19. Haas, W.; Faherty, B. K.; Gerber, S. A.; Elias, J. E.; Beausoleil, S. A.; Bakalarski, C. E.; Li, X.; Villen, J.; Gygi, S. P., Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Mol Cell Proteomics* **2006,** 5, (7), 1326-37.
20. Eng, J.; McCormack, A.; Yates, J., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of The American Society for Mass Spectrometry* **1994,** 5, (11), 976-989.
21. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007,** 4, (3), 207-14.
22. Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P., A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* **2006,** 24, (10), 1285-92.
23. Kim, W.; Bennett, E. J.; Huttlin, E. L.; Guo, A.; Li, J.; Possemato, A.; Sowa, M. E.; Rad, R.; Rush, J.; Comb, M. J.; Harper, J. W.; Gygi, S. P., Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol Cell* **2011,** 44, (2), 325-40.
24. Schwartz, D.; Gygi, S. P., An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol* **2005,** 23, (11), 1391-8.
25. Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E., WebLogo: a sequence logo generator. *Genome Res* **2004,** 14, (6), 1188-90.
26. Khidekel, N.; Ficarro, S. B.; Clark, P. M.; Bryan, M. C.; Swaney, D. L.; Rexach, J. E.; Sun, Y. E.; Coon, J. J.; Peters, E. C.; Hsieh-Wilson, L. C., Probing the dynamics of O-GlcNAc glycosylation in the brain using quantitative proteomics. *Nat Chem Biol* **2007,** 3, (6), 339-48.
27. Shahbazian, D.; Roux, P. P.; Mieulet, V.; Cohen, M. S.; Raught, B.; Taunton, J.; Hershey, J. W.; Blenis, J.; Pende, M.; Sonenberg, N., The mTOR/PI3K and MAPK pathways converge on eIF4B to control its phosphorylation and activity. *EMBO J* **2006,** 25, (12), 2781-91.

28. Jensen, C. J.; Buch, M. B.; Krag, T. O.; Hemmings, B. A.; Gammeltoft, S.; Frodin, M., 90-kDa ribosomal S6 kinase is phosphorylated and activated by 3-phosphoinositide-dependent protein kinase-1. *J Biol Chem* **1999,** 274, (38), 27168-76.
29. Payne, D. M.; Rossomando, A. J.; Martino, P.; Erickson, A. K.; Her, J. H.; Shabanowitz, J.; Hunt, D. F.; Weber, M. J.; Sturgill, T. W., Identification of the regulatory phosphorylation sites in pp42/mitogen-activated protein kinase (MAP kinase). *EMBO J* **1991,** 10, (4), 885-92.
30. Fuentes, J. J.; Genesca, L.; Kingsbury, T. J.; Cunningham, K. W.; Perez-Riba, M.; Estivill, X.; de la Luna, S., DSCR1, overexpressed in Down syndrome, is an inhibitor of calcineurin-mediated signaling pathways. *Hum Mol Genet* **2000,** 9, (11), 1681-90.
31. Genesca, L.; Aubareda, A.; Fuentes, J. J.; Estivill, X.; De La Luna, S.; Perez-Riba, M., Phosphorylation of calcipressin 1 increases its ability to inhibit calcineurin and decreases calcipressin half-life. *Biochem J* **2003,** 374, (Pt 2), 567-75.
32. Wang, Y.; Li, G.; Goode, J.; Paz, J. C.; Ouyang, K.; Screaton, R.; Fischer, W. H.; Chen, J.; Tabas, I.; Montminy, M., Inositol-1,4,5-trisphosphate receptor regulates hepatic gluconeogenesis in fasting and diabetes. *Nature* **2012,** 485, (7396), 128-32.
33. Burchfield, J. G.; Lennard, A. J.; Narasimhan, S.; Hughes, W. E.; Wasinger, V. C.; Corthals, G. L.; Okuda, T.; Kondoh, H.; Biden, T. J.; Schmitz-Peiffer, C., Akt mediates insulin-stimulated phosphorylation of Ndrg2: evidence for cross-talk with protein kinase C theta. *J Biol Chem* **2004,** 279, (18), 18623-32.
34. Joo, H. Y.; Zhai, L.; Yang, C.; Nie, S.; Erdjument-Bromage, H.; Tempst, P.; Chang, C.; Wang, H., Regulation of cell cycle progression and gene expression by H2A deubiquitination. *Nature* **2007,** 449, (7165), 1068-72.
35. Poulin, F.; Gingras, A. C.; Olsen, H.; Chevalier, S.; Sonenberg, N., 4E-BP3, a new member of the eukaryotic initiation factor 4E-binding protein family. *J Biol Chem* **1998,** 273, (22), 14002-7.
36. Yang, H. S.; Jansen, A. P.; Komar, A. A.; Zheng, X.; Merrick, W. C.; Costes, S.; Lockett, S. J.; Sonenberg, N.; Colburn, N. H., The transformation suppressor Pdcd4 is a novel eukaryotic translation initiation factor 4A binding protein that inhibits translation. *Mol Cell Biol* **2003,** 23, (1), 26-37.

Chapter 4

Proteome-Wide Applications of Quantitative Multiplexing in the Yeast Stress Response

Attributions:

- J.T. Wilson-Grady performed all experimental steps and data analysis. All bioinformatic processing was performed using in-house software developed by the “GFY Development Team.”
- D.P. Nusinow generated PCA and NMF data, the associated plots, and the residual plots.
- S.P. Gygi advised the project

Abstract

The use of LC-MS/MS for peptide identification and quantification has become the preferred method for proteome-wide analyses. Traditionally in LC-MS/MS based quantification, the relative abundance of a peptide and its stable isotope enriched pair (e.g. ^{13}C, through chemical or metabolic labeling) are used to obtain a ratio which reflects the relative expression of a protein under the conditions tested. Though MS^1 based quantitative methods such as SILAC have been successfully applied in numerous analyses, they also contain several drawbacks. First, in MS^1 based quantification, a peptide's quantification is tied to its identification, which may lead to erroneous quantifications when proper false discovery rate filters are not imposed (e.g. false positive hits cluster in the group of peptides which change by two-fold or greater). Second, due to the stochastic nature of shotgun proteomics, combining multiple peptide labels in one LC-MS/MS analysis and combining multiple experiments into one data set lead to decreased proteome coverage (due to increased sample complexity) and missing quantitative values between experiments, respectively. MS^2 based quantitative techniques, such as peptide labeling with tandem mass tags (TMT), avoid theses drawbacks and permit unique proteome wide analyses.

Three common experiments which are made possible by TMT are demonstrated in this chapter and include replicate analysis for true statistical comparisons, time course analysis, and discovery based approaches (multi-state comparisons). These experiment types are demonstrated in the context of the yeast environmental stress response. This system was chosen as it is likely that protein-level regulation occurs in the stress response (e.g. protein degradation), it may be applicable to human diseases such as Alzheimer's disease and cancer. These data sets will serve to complement the wealth of available genetic data on the yeast stress response, which may shed light on transcriptional vs. translational or protein-level regulatory mechanisms.

Introduction

From Genomics to Proteomics

One can draw several parallels between the field of genomics and the field of proteomics. The goal of both fields is to identify and quantify their respective biomolecules in a high throughput manner, thus generating the type of large scale data that is required to answer many of the complex questions that encompass biology. As the field of proteomics continues to progress, we see that not only does it share similar goals as genomics, but also shares a common evolution. Analogous advancements that paved the way for modern genetic analyses are being made in the field of mass spectrometry based proteomics, thus providing insight into the direction we must go in order to fully appreciate the promise of the field.

The discipline of genomics began with Fredrick Sanger's invention of the "plus and minus end" sequencing technique[1], which was laborious, costly and slow, yet still produced the first genome, that of the bacteriophage phi X174[2]. Despite its limitations, this method is widely regarded as a milestone in biological research. This technique was further developed with the invention of the chain terminator method[3], which increased the speed and coverage of sequencing, while reducing the cost. Further developments in speed, sensitivity, and multiplexing, through new reagents (e.g. dye terminators), methods (e.g. multiplex tagging[4]) and instrumentation (e.g. ABI 370) permitted the sequencing of genomes at an unprecedented rate (indeed we may be in the "personal genomics era")[5]. Similar developments in quantitative genomics gave rise to accurate genome-wide microarrays, which can tolerate the analysis of multiple samples simultaneously (e.g. multi-color SNP microarrays[6]). In a similar manner, the field of proteomics has evolved from a low throughput qualitative endeavor to a nearly comprehensive quantitative technology. High accuracy and fast scanning mass spectrometers[7], and

quantitative techniques such as SILAC[8] and reductive dimethylation, are just a couple of examples of these developments.

The Rise of Quantitative Multiplexing, Technical Hurdles and Solutions

The general trend we observe in these fields is that new methods and technology development - which increase sensitivity, throughput, and permit multiplexing - further the usefulness of these disciplines to answer relevant biological questions. Until recently, mass spectrometry based proteomics lagged behind genomics in its ability to be comprehensive and to offer greater than binary comparisons. However, with the development new technologies surrounding the use of tandem mass tags (TMT) and Isobaric tag for relative and absolute quantitation (iTRAQ), we now have the potential for proteome-wide quantitative multiplexing. The promise of large scale multiplexing is enormous in both basic and clinical research, and the need to evaluate the quality of available methods and demonstrate their applications is great. Once proven as a valuable technique, multiplexing will allow for deeper biological inquiry, and permit us to answer some unique biological questions.

Currently six versions of TMT exist commercially (only TMT is used in this chapter, structures are summarized in Figure 4.1, referred to as channels 126-131), which enable the simultaneous quantitative comparison of six biological states. The chemistry of the reaction involves an NHS-ester addition of the reagent to the free amines of lysine residues and peptide N-termini. The reagents are isobaric, meaning each reagent adds the identical mass to peptides in all cases. As visible in the structure, the isobaric nature of the reagents comes from the distribution of ^{13}C and ^{15}N on either side of a cleavage site (dashed line in Figure 4.1), which is required for differential quantification of each biological state. Unlike MS^1 based quantification methods, such as SILAC, TMT quantification occurs in the MS^2 spectrum after peptide fragmentation by higher-energy collision dissociation (HCD). Fragmentation and quantification is also possible by CID using quadrupole time-of-flight (Q-TOF) mass spectrometry, though

these instruments cannot currently match the analytical depth of the LTQ-Orbitrap-Velos. During fragmentation, the bond at the cleavage site is broken, and reporter ions (left side of the molecule, with respect to the cleave site, are liberated. The benefits of such a quantification scheme are many.

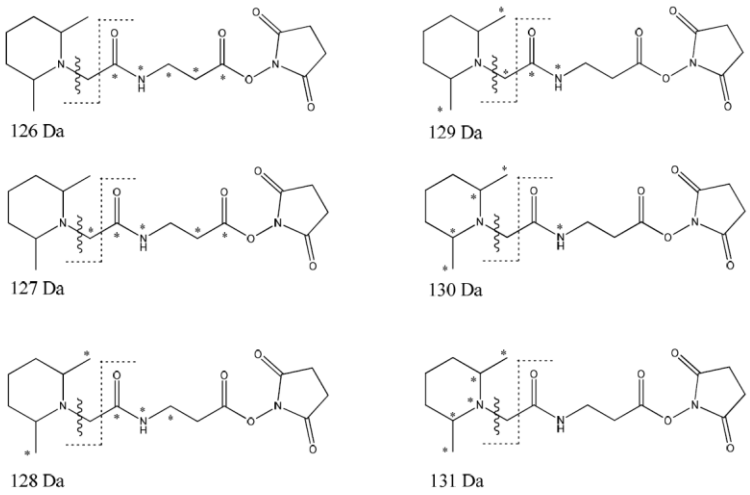


Figure 4.1. The structure of the Tandem Mass Tag (TMT) reagents permits quantitative multiplexing. Up to six samples may be labeled with the six different varieties of TMT reagent, which modify free amines (lysine and N-terminus) and are denoted henceforth as 126, 127, 128, 129, 130 and 131 (for their approximate reporter ion mass). Each reagent contains isotopomerically related reporter ion and linker regions (separated by the dashed line), and are chemically identical. Heavy isotopes of ^{13}C and ^{15}N (denoted by the *) are distributed amongst the linker and reporter ion components of the reagent so that the intact regents are isobaric (each reagent modifies a peptide by 229.1629 amu/label). The reporter ions generated during MS/MS are distinct. For example the 126 reagent contains no isotopes in the reporter ion region, whereas the 131 reagent contains five. Figure adapted from manufacturer's instructions (http://www.piercenet.com/instructions/2162073.pdf).

The first obvious benefit of TMT is that many samples are simultaneously analyzed, which reduces preparation and analytical time. Due to the stochastic nature of LC-MS based proteomics, the analysis of complex proteomes often leads to a unique set of identified peptides between experiments, even when the identical samples are analyzed. The result of this phenomenon is the presence of missing values between quantitative data sets (such as by SILAC), limiting their comparative potential. In contrast, TMT avoids missing values between experimental conditions, within each set of 6 biological states. The described isobaric nature of TMT also allows these samples to be combined without increasing proteome complexity (Figure 4.2). In the MS^1 spectrum, peptides from all six biological states

will be observed at the same *m/z* (mass to charge ratio), and only upon MS^2 fragmentation are the unique reporter ions generated, from which ratio between each state are obtained. Peptide identifications are also obtained in this step, as standard b- and y-type ions are generated alongside the low mass reporter ions. A drawback of methods such as SILAC is that each additional isotopic form of a peptide (light, medium, heavy, etc.) increases sample complexity, thus reducing unique peptide and protein identification. One does not affect peptide identification through the use of TMT in this manner. Though TMT demonstrates many useful properties, it contains its own glaring drawback which has to date limited it use, the presence of interference.

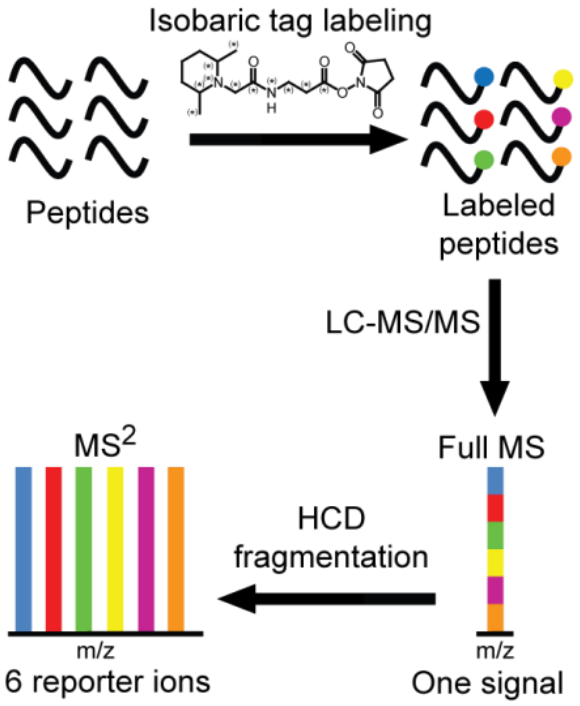


Figure 4.2. TMT reagents allow quantitative multiplexing, without increasing sample complexity. Due to the isobaric nature of the reagents, multiple samples can be labeled and combined without increased proteome complexity. As with other stable isotope methods relying on ^{13}C and ^{15}N incorporation, the TMT reagents are chemically identical are indistinguishable by chromatographic separation. Additionally, in a full MS spectrum, a peptide labeled with any of the six TMT reagents will have the same mass to charge ratio (m/z), thus maintain the complexity of an unlabeled samples. This behavior contrasts the increase in sample complexity observed with MS^1 quantification methods, such as SILAC and ReDi. Only once TMT labeled peptides are fragment by higher energy collisional dissociation (HCD), are the reporter ions generated. Peptide ratios (and thus protein ratios) are determined by the S/N ratios of the reporter ions. Figure adapted from *MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics*[10] , used with permission.

Interference is the unintended co-isolation of contaminating TMT labeled peptides in the MS^1 spectrum along with the peptide of interest (Figure 4.3), prior to MS^2 fragmentation. The observed reporter ion signal in the MS^2 is an amalgamation of intended and unintended peptides. The result of this contamination is the compression of observed ratios among the six TMT channels, as the majority of

peptides contained within an LC-MS run are typically observed at 1:1 ratios with respect to all biological conditions. As such the identification of proteins which are up and downregulated in an experiment is hindered. Interference is unavoidable and cannot be abrogated through increased peptide fractionation, or decreased MS^1 isolation windows.

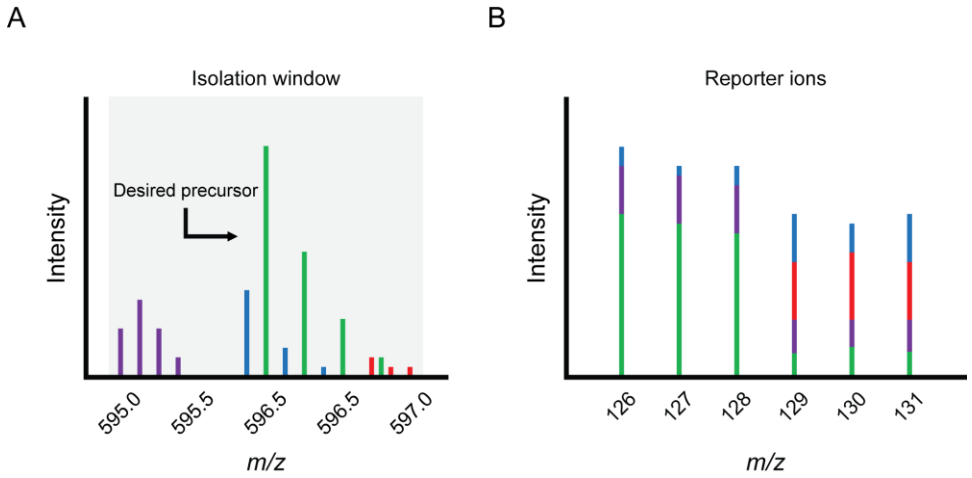


Figure 4.3. Interference compresses observed ratios, limiting the accuracy and biological relevance of TMT data. (A) While isolating a peak in an MS^1 spectrum for MS^2 analysis, unintended MS^1 ions (purple, blue and red peaks) are often isolated along with the desired precursors ions (green peaks), thus creating a in impure MS^2 spectrum which contains peptide fragment ions and reporter ions from unwanted contaminants. (B) The result of this contamination is a compression of the true ratios among the reporter ion channels (green peaks) toward 1:1 ratios (all peaks), as most reporter ions will on average be close to a 1:1 ratio among the six channels. Thus a ratio that may actually be 5:1 could be reported at 1.2:1, depending on the exact nature of the interference.

The abrogation of interference is a subject of intense research and has been achieved, for example, by the gas phase isolation of the charge reduced species (formed during ETD fragmentation) corresponding to the ion of interest prior to HCD fragmentation[9]. This method however does not completely eliminate interference and is very costly to the MS duty cycle. A means of eliminating interference is through gas phase isolation of intended ions in an MS^2 spectrum (Figure 4.4) from contaminating ions in the linear ion trap. In such a method (referred to as the MS^3 method), an MS^1 ion of interest is isolated for MS^2 fragmentation by collision induced dissociation (CID), along with the discussed contaminating ions. Identification of peptides still occurs using this MS^2 spectrum. In the MS^2 spectrum, one of the fragment ions is re-isolated (above the m/z range of the parent ion to avoid re-

isolation of contaminants) for fragmentation by HCD, and quantitation of the TMT reporter ions. Thus the identification and quantification of a peptide is decoupled, which has been proven to functionally avoid interference[10]. A limitation of this method, as will be elaborated upon, is that the abrogation of interference comes at a price of TMT signal loss, which in many cases limits the ability to accurately quantify peptides. In this chapter, I evaluate technological advancements which have made the wide spread use of TMT possible.

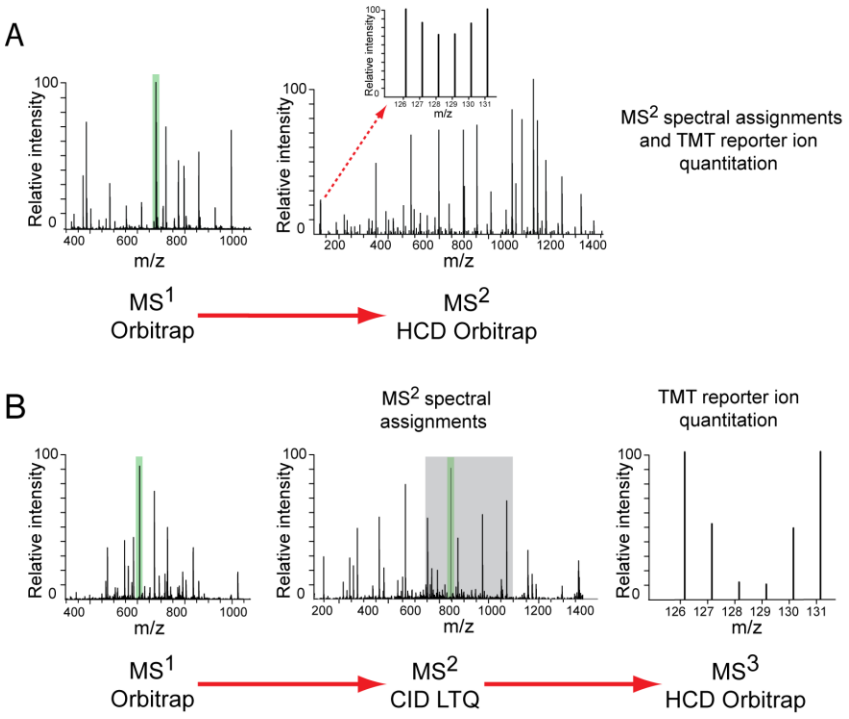


Figure 4.4. A comparison of the MS^2 and MS^3 methods for peptide identification and quantification. (A) Using the MS^2 method, peptide identification and reporter ion quantification occur simultaneously by HCD. This method leads to the co-isolation and fragmentation of unwanted ions which compresses the observed TMT ratios. (B) The MS^3 method decouples the peptide identification and reporter ion quantification steps. Peptide fragmentation occurs in the MS^2 spectra by CID, generating b- and y-ions, from which a peptide's identification can be ascertained. This fragmentation step, however does not liberate the TMT reporter ions, which remain on the b- (and lysine contain y-) ions. An MS^2 ion is then isolated, from a region above the parent ion m/z, to avoid re-isolation of contaminants, for fragmentation by HCD (generating an MS^3 spectrum). The peptide from all samples is quantified using the reporter ions in the MS^3 spectrum. This gas phase purification step abrogates interference. In each step the, the intended ion for isolation is highlighted in green. Figure adapted from *MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics*, used with permission.

Applications of TMT to Biological Inquiry

After a discussion on technological improvements that make the use of TMT applicable to biological inquiry, I demonstrate the use of TMT in the three types of common proteomics experiments: statistical analysis of biological replicates, time-course measurements, and discovery experiments (multi-state comparisons). Statistics have become a requirement for modern biological analyses. Often triplicate analysis of a given biological response is sufficient to apply relevant statistical tests, which assign significance to an observation. With statistics, smaller changes may become more relevant and useful in unraveling the underlying biology. With the power quantitative multiplexing, we have the ability to dissect both the obvious large changing proteins, and also the more subtle differences, by applying statistics on a proteome-wide scale. In a time course analysis, on the other hand, though no replicates exist (within one LC-MS experiment), each successive time point acts as a pseudo replicate. The overall pattern of protein expression allows one to classify each protein into temporal groups. Stable protein expression patterns vs. stochastic variability over time allow one to determine which proteins may be relevant throughout the time course. Commonalities among the proteins within each temporal group will likely be relevant to the biology at hand. Finally, in a similar manner to the time course experiment, similarities and differences among the many conditions in a discovery experiment are useful for extracting relevant proteins from the background of stochastic biology. The depth of the presented analyses is such that the discussed patterns can indeed be found. This systems biology approach enables a greater analytical depth to be obtained, far beyond that of the simple distribution statistics which has until now been commonly applied to quantitative LC-MS data sets. A common theme among all experiments is the need to handle complex data sets; for such analyses, understanding the relationships amongst the experimental conditions and amongst the proteins themselves, and a reduction of variables to a manageable number of components is paramount.

Bioinformatic Tools for Interpreting Complex Data Sets

Common means of interpreting large scale data sets include cluster analysis (hierarchical and K-means) and principal component analysis (PCA). In hierarchical clustering, a relationship of decreasingly related proteins and samples is constructed (each separately) based on the protein expression values in the data matrix. The generated dendrogram is useful for identifying related groups which may have a biological importance. In K-means clustering, proteins are partitioned into a pre-defined (by the user) number of groups, usually based on the biology at hand. In a manner similar to hierarchical clustering, these groups may contain proteins which have related biology, though K-means does not identify relationships among data points. In conjunction with clustering methods, PCA is a useful method for large scale data interpretation. Similar to K-means, PCA is useful for reducing multivariate data into a manageable number of components. Often the samples within an analysis achieve separation among different components, which is useful for defining which protein features are shared and which are unique among the conditions tested. Unlike K-means, however, PCA defines components based on observed variance in the data (each component explains a fraction of total variance), independent of the biology at hand. As such, explaining the relevance of each component is often a non-trivial endeavor. A novel method of large data set reduction, non-negative matrix factorization (NMF), has recently proved to be useful in biology, particularly for the analysis of microarray data sets[11, 12]. NMF in a way can combine aspects K-means and PCA, in that a desired number of clusters (based on the biology) can be chosen, and the original matrix of protein expression is deconvoluted through its factorization into many matrices (variable reduction). In contrast to PCA, clusters discovered through NMF are often readily interpretable. As discussed later, NMF has additional properties which are useful for proteomic analyses.

Demonstrating the Capabilities of Proteome-Wide Multiplexing in the Yeast Stress Response

I chose to demonstrate the discussed experiment types and analytical techniques within the context of the yeast stress response. This system is ideal for proteomic demonstrations: though genetic studies on the subject are available, many which and have established important genomic responses, proteome level data is lacking. Hence, the yeast stress response has a solid biological framework within which proteomics data may be interpreted, though the protein level data itself embodies novel aspects of the system. From a biological perspective, the stress adaptation mechanisms themselves are also of interest. In contrast to stress adaptation in mammals, where a near constant internal environment is maintained by numerous hemostatic processes, unicellular yeast face a changing and often hostile environment. As a result, they have evolved numerous adaptation strategies for coping with variations in temperature, salt concentration, nutrient availability, the presence of toxins, and other factors. Such environmental factors require yeast to maintain proper protein folding, redox conditions, protein turnover, membrane dynamics, metabolic homeostasis, DNA fidelity, etc. Based on the plethora of stress related processes, it is likely a variety of proteins are regulated during stress adaptation.

The primary purpose of these experiments is to demonstrate the robust nature of our strategy for proteome wide quantitative multiplexing, and to highlight surrounding analyses, in order to facilitate future experimentation. With this in mind, much time is given to the analytical techniques and data analysis. However, the quality of the data generated also presents an opportunity to remark on protein level adaptation in the yeast stress response. As such many relevant proteins and pathways are highlighted, and they are understood in relation to published literature. As such, these data are a valuable resource to the yeast community.

Material and Methods

Yeast Strains and Culture conditions

In all cases the BY4742 strain of Wt yeast (MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0) was used. Yeast were grown in synthetic complete media, containing all amino acids. Three experiments were performed. An experiment comparing five stress conditions to an unstressed control was first undertaken. The chosen stress conditions were cold (media pre-cooled to 10 °C), oxidative (0.3 mM H_2O_2 in the media), osmotic (0.7 M NaCl in the media), heat (media pre-heated to 37 °C) and cytotoxic/endoplasmic reticulum (ER, arginine replaced by 76 mg/L canavanine in the media) stresses. A large overnight culture of unstressed yeast was grown to mid log phase at 30 °C. This culture was divided into six tubes (one for each condition), gently pelleted (3000 rpm), and the media was removed. Pellets were resuspended in the appropriate media for each stress condition. Controls were resuspended in 30 °C media without additives. The canavanine treated yeast were pelleted (media removed) once more and resuspened in the appropriate media; this step was undertaken to ensure the canavanine stressed culture was free of remaining arginine. The stresses and control cultures were grown for 1hr at 30 °C, with the exception of the cold and heat stresses which were grown at 10 °C and 37 °C, respectively. This experiment was also repeated with a 2 hr time point. A heat stress time course was performed in a similar manner, though overnight cultures were grown at 25 °C. As before, yeast were stressed at 37 °C, and 0, 30, 60, 90, 120, and 240 minute time points were collected. Finally, a triplicate analysis of heat stress was undertaken. Three separate overnight cultures were grown at 25 °C, and were stressed at 37 °C as before. 0 and 60 minute time points were collected for each replicate. ~20 OD_{600} units per condition were harvested in all cases, combined with an equal volume of 40 mM NaN_3 in ice cold dH_2O (to inhibit further biology) and pelleted (5, 000 RPM, 10 min). The pellets were

washed with ice cold dH_2O and frozen in liquid nitrogen. All harvested yeast samples were in the logarithmic growth phase.

Cell lysate preparation

A lysis buffer containing 8M urea, 75mM NaCl, 50 mM HEPES pH 8.8, and a protease inhibitor cocktail (2 tablets of Roche complete mini per 10 mL and 1mM PMSF) was prepared. The ~20 OD_{600} of pelleted cells from each condition were resuspended in 1 mL of ice cold lysis buffer and transferred to a 2mL screw cap microcentrifuge tube containing ~1ml of 0.5 mm silica beads. Cells were lysed by bead beating at maximal power (Mini-Bead Beater 8, Biospec) for 3 pulses of 45 seconds, at 4 °C. Tubes were cooled on ice between homogenization cycles to prevent protein degradation. The lysates were separated from the beads, and insoluble components were removed by centrifugation (14,000 rpm, 5 min at 4°C). Protein concentrations were determined using the BCA method in triplicate.

Reduction, alkylation, precipitation, and digestion of proteins

100 μg of protein were used for each condition, and all volumes were equalized within one experiment by the addition of lysis buffer when required; equal volumes simplified downstream steps. Disulfide bonds were reduced in 5mM DTT at 56 °C for 45 min. Free sulfhydryl groups were then alkylated in 15mM iodoacetamide, in the dark at room temperature for 45 min. The reaction was quenched for 15 min at room temperature, in the dark with another addition of DTT, to the final concentration of 5 mM.

Proteins were precipitated using methanol-chloroform extraction: 4x volumes of MeOH, 1x volume of chloroform and 3x volume of dH_2O were added to the samples, vortexing between each addition. The mixtures were centrifuged at room temperature for 5 min (6000 rpm). The aqueous layer (above white protein pellet) was removed. A 4x volume of MeOH was again added to each sample, and

each was vortexed. Samples were centrifuged again for 5 min (6000 RPM), the supernatant was removed, and the tubes were air dried for 10 min. Protein pellets remained at the bottom of the tube.

Each pellet was resuspended in 100 µL of 8M urea in 50 mM HEPES pH 8.8. If required, the pellets were bath sonicated to facilitate protein resuspension. Samples were diluted to 2M urea with 50 mM HEPES pH 8.8. The endopeptidase Lys-C (cleaves after lysine residues) was added at a 1:50, enzyme to substrate ratio (2 µg of enzyme per 100 µg protein). Digestion occurred overnight at room temperature. The digested peptides were acidified with formic acid and desalted on C18 Sep-Paks as discussed in chapter 3. Samples were dried to completion by vacuum centrifugation.

Peptide TMT labeling

TMT reagents (each containing 0.8 mg of lyophilized reagent) were resuspended in 40 µL of anhydrous acetonitrile (ACN). Each sample was resuspended in 100 µL of 200 mM HEPES pH 8.5 and 30 µL ACN. 10 µL of each resuspended reagent (126-131) was added to the proper sample, and vortexed. The samples were incubated at room temperature for one hour. The reaction was quenched by the addition of hydroxyl amine to the final concentration of 0.3% (vol/vol) for 15 min. The samples were combined at 1:1 ratios for all channels and dried by vacuum centrifugation. The combined sample pellet was desalted on C18 Sep-Paks as discussed in chapter 3.

In the five stresses experiment, the control was labeled with the 126 reagent, and the cold, oxidative, osmotic, heat, and cytotoxic/ER stress were labeled with the 127-131 reagents respectively. In the heat stress time course the control was again labeled with the 126 reagent, and the successive time points were labeled with the 127-131 reagents. In the triplicate analysis of heat stress, the controls were labeled with the 126-128 reagents and the heat stress samples were labeled with the 129-131 reagents.

Peptide pre-fractionation

Peptides were fractionated by high pH reverse phased chromatography (HPRP). The combined TMT labeled peptides were rehydrated in 500 µL of 10 mM ammonium formate/5% ACN pH 10. Buffer A was 10 mM ammonium formate/5% ACN pH 10 and buffer B was 10 mM ammonium formate/90% ACN pH 10. The peptides were separated over a one hour gradient of 5 to 25 % B. 96 fractions of 38 seconds were collected in a 96 plate. Fraction were pooled into 12 fractions as discussed in Figure 4.7, and desalted as discussed in chapter 3 and dried to completion by vacuum centrifugation.

In cases where SCX was used, the same protocol discussed in chapter 3 was used with the following exceptions: a 4.6 mm column and 1 mL/min flow rate were used instead of the 9.4 mm column and 3 mL/min flow rate. 600 µg was separated over a 1 hour gradient, and 20 one minute fractions were collected. The buffers and gradient were the same. These fractions were desalted as discussed in chapter 3 and dried to completion by vacuum centrifugation.

LC-MS analysis

Dried peptides were resuspended in 10 µL of 5% FA/4% ACN. 2 µL of each sample was analyzed on an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific) equipped with an Accela 600 quaternary pump (Thermo Fisher Scientific) and a Famos Microautosampler (LC Packings). Nanospray tips were hand-pulled with 100-µm (inner diameter) fused-silica tubing and packed with 0.5 cm of Magic C4 resin (5 µm, 100 Å; Michrom Bioresources) followed by 20 cm of Maccel C18AQ resin (3 µm, 200 Å; Nest Group). Buffer A was 0.125% FA in 3% ACN and Buffer B was 0.125% FA in 100% ACN. Peptides were separated over a 3hr gradient from 5 to 28% buffer B, at a flow rate of ~500 nl/min. Peptides were detected in a hybrid dual-cell quadrupole linear ion trap–orbitrap mass spectrometer (LTQ Orbitrap Velos, Thermo Fisher) using a data-dependent Top10 MS^2/multinotch MS^3 method[13] (Figure 4.10). In each cycle, one full MS scan of mass/charge ratio (*m/z*) = 300 to 1500 was acquired in the Orbitrap at a

resolution of 60,000 at m/z = 400 with an automatic gain control (AGC) target of 2 x 10^6. Maximum injection time was set to 1000 ms for MS^1 scans.

Each full MS scan was followed by the selection of the top 10 most intense ions for collision-induced dissociation (CID) in the linear ion traps, for peptide identification. Higher-energy collisional dissociation (HCD) was subsequently used after the MS^2 scan for multinotch MS^3 analysis in the Orbitrap (7, 500 resolution), to quantify TMT reporter ions. AGC targets of 2 x 10^3 and 1 x 10^5 were used for MS^2 and MS^3 scans, respectively. Maximum injection times of 150 and 250 ms were used for MS^2 and MS^3 scans, respectively. Ions selected for MS^2 analysis were excluded from reanalysis for 120 s. In this new multinotch MS^3 method, multiple MS^2 fragment ions were captured in the MS^3 precursor population, using isolation waveforms with multiple frequency notches[13]. For such a method, online algorithms that filter MS^2 fragment ions based upon their expected reporter ion fragment signal and their required isolation specificity are used. In general 6–9 MS^2 fragment ions contributed to the MS^3 spectrum. In cases where the multinotch method was not used, a single MS^2 ion (at 110%-160% of the precursor ion *m/z*) was isolated for the MS^3 spectrum as discussed in Figure 4.4.

Database searching and filtering

RAW files obtained from data collection were converted into mzXML format using the ReAdW program (http://sourceforge.net/projects/sashimi/files/ReAdW%20%28Xcalibur%20converter%29/). MS/MS spectra were searched using SEQUEST v.28 (rev. 13) against a composite database containing the all predicted open reading frames of S. cerevisiae (http://downloads.yeastgenome.org, downloaded 30 October 2009) in their forward and reversed orientation. The following search parameters were used: a precursor mass tolerance of ±25 parts per million (ppm), a 1.0 Dalton fragment ion mass tolerance, LysC digestion specificity, and up to two missed cleavages. Static modifications of carbamidomethylation on cysteine (+57.0214), and TMT reagent additions (+229.1629) on lysine

residues and peptide N-termini were included. The dynamic modification of methionine oxidation (+15.9949) was allowed.

Matched peptide spectra were first filtered using a target-decoy strategy[14] to a 1% peptide level false discovery rate (FDR) through linear discriminant analysis (LDA) using the following parameters: XCorr, ΔCn', precursor mass error, solution charge (when analyzing SCX fractions only), observed ion charge state, and number of missed cleavages[15]. Linear discriminant models were calculated for each run using peptide matches to forward and reversed protein sequences as training data. Peptides contained within each MS/MS run were ranked by descending discriminant score and filtered to a 1% FDR based on the number of reverse sequences remaining (FDR = 2 * number of reverse hit/total hits). The data was subsequently filtered to control the protein level FDR. Proteins were scored by multiplying peptide probabilities, sorted by rank, and filtered to 1% FDR as described for the above peptides[15].

Peptide quantification and protein assembly

Reporter ion signal to noise (S/N) ratios were extracted for each MS^3 scan from the mzXML files, and were used for peptide quantification, with in-house software. Each reporter ion measurement was corrected for isotopic overlap between reporter ions based on the manufacture's specifications. Generally a minimum summed (across all channels of a peptide) S/N filter of 100 was implemented in the analysis of final data sets. With this filter in place, no additional, channel-specific filters were required. All peptides were collapsed into proteins so that the minimum number of proteins required to explain all peptide observations remained. Peptide sequences which were common to multiple proteins were assigned to the protein with the largest number of peptide spectral matches. Protein quantification occurred through the summation of peptide TMT values; the sum of the S/N from for each peptide in a channel was equal to the final protein S/N for that channel. This method of protein

quantification is a type of weighted average and is based on the observation that quantification events with greater S/N are often more accurate.

Generally all peptide and protein levels values are discussed as the normalized (relative) S/N or intensity (functionally equivalent terms). These terms denote the amount of total TMT signal from a peptide (or collapsed protein) measurement, which is contained within a given channel; this values across all channels will sum to 100%. In some cases ratios (or $\log_2$ ratios) are discussed; in these cases they signify the stress condition/control ratio. When analyzing isoform specific data, all protein measurements were confirmed using only unique peptides, to avoid fallacious assignment of common peptides to the wrong isoform. In these cases, no errors were found.

Statistical analysis of heat stress biological triplicates

P-values were calculated using a two-tailed, unpaired T-test, allowing for heteroscedastic variance (Welsh's T-test). The normalized TMT S/N from the control samples comprised one data array, and the other array was composed of the normalized S/N values of the heat stressed samples. The T-test was performed after protein collapsing, using the summed TMT S/N (by channel) values. T-tests were corrected using the Benjamini-Hochberg method for multiple hypothesis testing[16].

Hierarchical clustering

Clustering was performed using the Cluster 3.0[17] program (downloaded from: http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm). The Euclidian distance metric and centroid linkage clustering method were used. Heat maps were visualized using the Java TreeView[17] program (downloaded from: http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm). Clustering was generally performed using the normalized TMT S/N values, except where noted. Groups of proteins were chosen visually, based on the dendrogram tree branches.

Principal component (PCA) and non-negative matrix factorization (NMF) analyses

Principal components analysis was performed using R (The R Core Team, 2012) and results were plotted using the ggplot2 package (http://ggplot2.org/). Non-negative matrix factorization (NMF) was performed with the NMF R Package[18] using the Brunet algorithm[12]. To estimate basis number and accuracy, 200 replicates using random samples were done for basis numbers 2-6. Accuracy across the different bases was estimated using the cophenetic correlation score and residual sum of squares. Cophenetic correlation measure clustering stability, and the residual sum of squares measures errors in matrix factorization. Observed results were compared to those from a random matrix to estimate over fitting, and observed cophenetic scores were much higher at all ranks, and residual sum of square was lower at all ranks. Ranks were chosen for this analysis because of a suitably high cophenetic correlation score and ready biological interpretation. Features were extracted using the score defined by Kim and Park[11]. Plotting was done with the NMF, pheatmap, and limma[19] packages.

Gene ontology (GO) analysis

In all experiments, the identification of significantly enriched GO categories was achieved using the DAVID Bioinformatics resource (version 6.7[20]). The list of foreground proteins is specified when discussing a particular analysis. The background for each analysis was the respective list of all quantified proteins within one experiment. All categories were required to pass a p-value cutoff (after Benjamini-Hochberg correction) of 0.05, unless specified, generally where biological interpretation was readily available.

Assembly of protein interaction networks using Genemania

GeneMANIA (http://www.genemania.org/)[21] was used to assemble protein-protein interaction networks (based on literature annotated physical interactions), using the Cytoscape plugin[22]. The

"automatically selected weighting method" was used to properly assign the top 40 interacting partners for the queried proteins. The queried proteins are described during the presentation of each network. Network figures were constructed in Cytoscape, using protein expression data generated in the experiments described above. Nodes represent proteins in the network, and edges represent a physical interaction. The thickness of an edge represents the confidence of interaction assigned by GeneMANIA. Nodes colored as follows: black, not quantified; gray, unchanged; blue, downregulated; red, upregulated. The GeneMANIA database was accessed in December 2012.

Results and Discussion

Experimental Design Overview

As discussed, the three major types of types of TMT comparisons are explored here: Discovery based (Figure 4.5, A), statistical comparisons of biological replicates (Figure 4.5, B), and time course measurements (Figure 4.5, C). Discovery experiments are particularly useful for identifying common and unique biological reaction to stimuli. The statistical element of biological triplicate analysis allows one to dig deeper into the data set for regulated proteins (those that change by small magnitudes), as a confidence score can be placed on all proteins. Time course measurements allow groups of proteins to be clustered based on their expression pattern, and these patterns often are indicative of regulation and biological function. Additionally, each time point in the series acts as a pseudo-replicate, in that the smoothness of the temporal data trend is indicative of proper quantification and biological relevance.

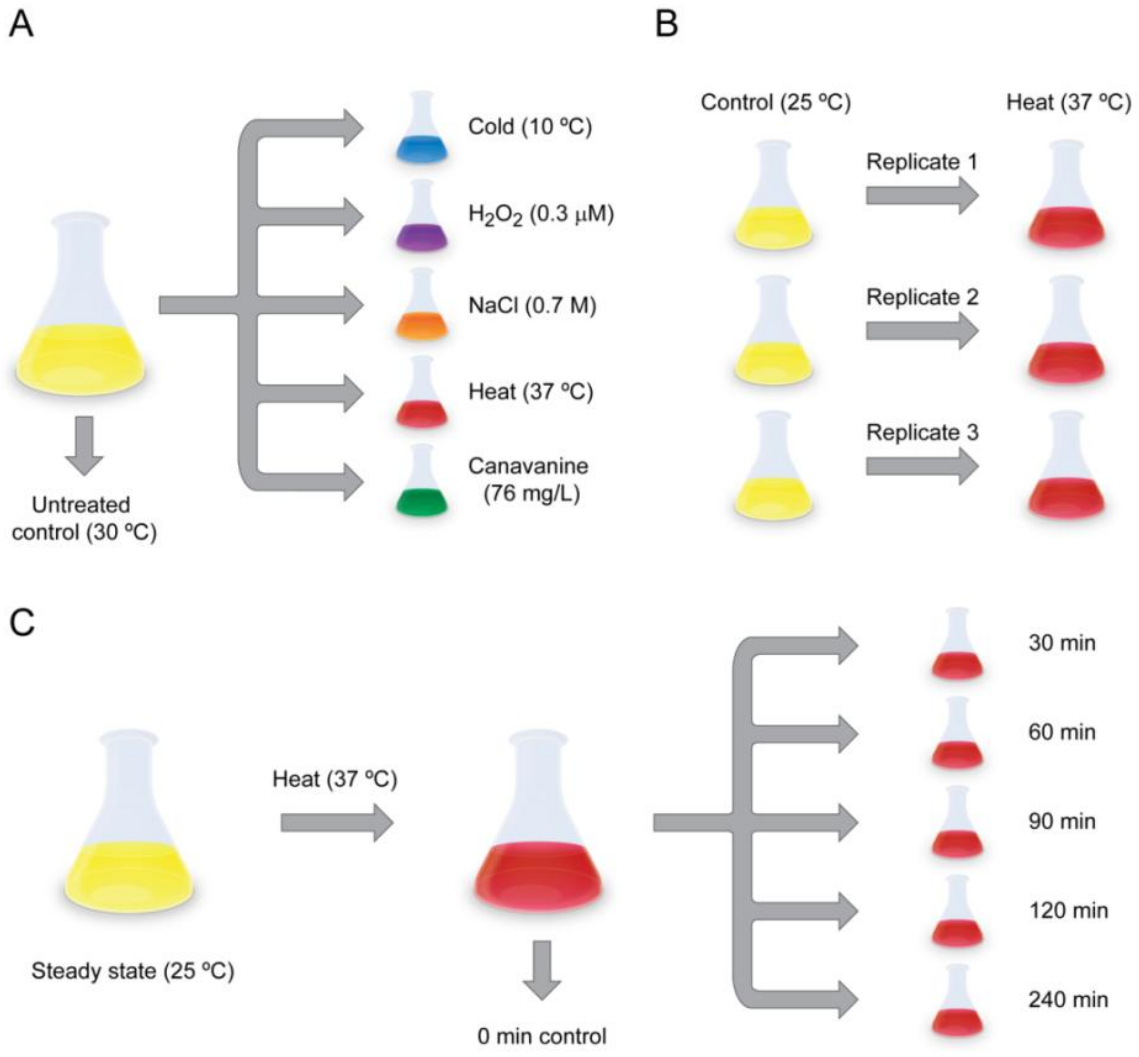


Figure 4.5. Experimental design of yeast stress experiments. Three common types of experiments made possible through multiplexing are demonstrated; Discovery (A, five stresses vs. an unstressed control), statistical comparisons (B, three biological replicates heat stress), and a time course (C, five heat stress time points vs. a 0 min control). In the discovery experiment, the control samples was labeled with the 126 reagent and the cold, H_2O_2, salt, heat and canavanine stresses were labeled with the 127-131 reagents, respectively. In the biological triplicate analysis of heat stress three control samples were labeled with the 126-128 reagents, and three heat stress samples were labeled with the 129-131 reagents. The 0 minute time point was labeled with the 126 in the time course experiment, and the successive time points were labeled with the 127-131 reagents. Each experiment is useful for answering different questions: discovery experiments in this case may reveal the common and unique stress responses in yeast. Replicate statistics, made possible in the biological triplicate analysis of heat stress, allow for the identification of more subtle (yet significant) changes in response to stress, such as small (~1.2 fold) change in protein abundance. Finally time course experiments allow for both the identification of sustained, delayed and transient regulation of proteins, as well increasing the change that a temporally regulated protein is identified (due to a larger number of data points).

In the discovery based analysis five yeast stress states were compared to an unstressed, steady state control. Cold (10 °C), oxidative (0.3 mM H_2O_2), osmotic (0.7 M NaCl), heat (37 °C) and cytotoxic/ER (76 mg/L canavanine in place of arginine in the media) stresses were explored. The comparison of yeast heat stress (37 °C) was chosen for both the biological replicate and time course analyses, to due the extensive available data in the literature on the subject. Generally each protein value in a given condition (e.g. 5 stress 128 channel, H_2O_2 treatment) is reported as its percentage of the total TMT signal for that protein (referred to as the normalized TMT intensity or normalized TMT signal/noise, S/N); thus the values for a protein in a given experiment sum to 100. This normalized reporting is particularly suited for large scale analytical methods, including PCA and hierarchical clustering, due to the inclusion of control samples (ratios eliminate the direct weight of the control). Unless otherwise specified, when referring to ratios (logged or unlogged), they represent the stress condition S/N divided by the control condition S/N for the five stress and time course data sets, and the sum of the heat stress replicate S/N divided by the sum of the control replicate S/N. All TMT channel values for a protein are obtained by summing that channel S/N from all peptides assigned to that protein (a type of weighted average).

In all experiments the data was obtained and processed though our in house proteomics pipeline (Figure 4.6). This pipeline includes protein isolation, digestion, TMT labeling, peptide fractionation, LC-MS analysis and database searching, false-discovery rate estimation and quantification (bioinformatics). A new approach for peptide pre-fractionation, named high pH reverse phased chromatography (HPRP), is discussed below. This novel method (in terms of its application in proteomics) facilitates greater proteome coverage. Additionally a novel method for TMT quantification, named “multinotch”, which improves the quality of quantification, is presented. Combined, these methods permit the use of TMT for true proteome wide quantitative comparisons in a multiplexed fashion.

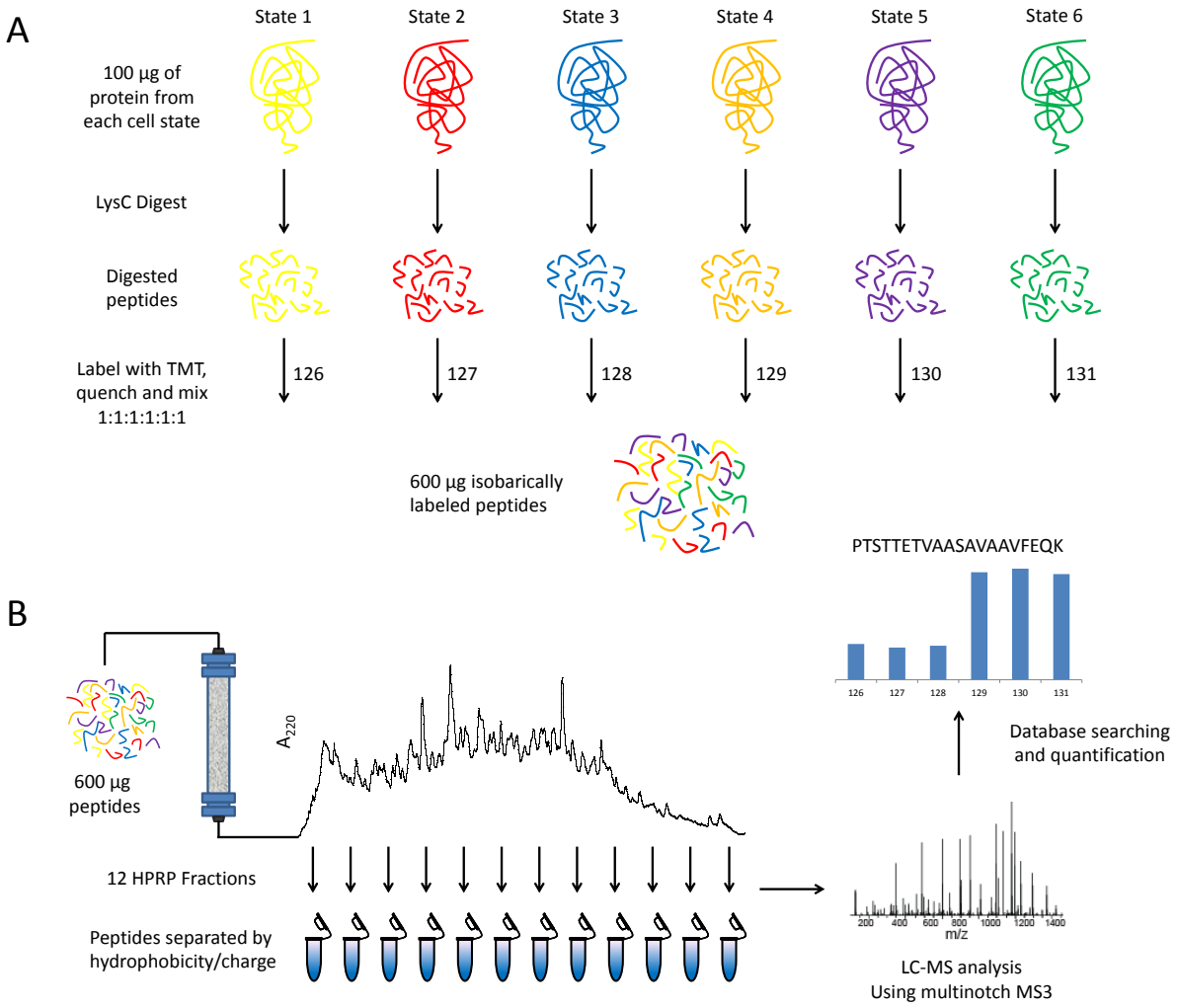


Figure 4.6. General protocol for proteome wide quantitative multiplexing. Commercially available TMT (Thermo scientific) allows for 6-plex quantification on a proteome wide scale (A). Samples are collected and lysates are prepared separately, as no metabolic labels are present. 100 µg of protein from each lysate are digested (in these experiments with LysC) prior to a reverse phased cleanup step. Desalted peptides are then labeled with the isobaric TMT regents (labeled 126-131, based on the reporter ion masses). The labeled peptides are then mixed at a 1:1 ratio for all samples. Mixed peptides are separated by high pH reverse phased chromatography (HPRP), prior to LC-MS/MS, and data base searching (B). Peptides are separated into 12 fractions based on their hydrophobicity at pH 10. Separated peptide samples are analyzed LC-MS/MS, using a multinotch MS^3 method for reporter ion quantification. Raw data from the LC-MS/MS is extracted and peptides are matched using the SEQUEST algorithm. Reporter ion signal to noise measurements are used to quantify between samples.

Technological Improvements Enabled True Proteome Wide Quantitative Multiplexing

Improvements in sample preparation and on-line analytical techniques have provided a framework for accurate, proteome-wide quantification using TMT.

Improvements in Peptide Pre-Fractionation

High pH reverse phased separation for peptide pre-fractionation is a novel and interesting method which demonstrates improvements over the current standard, strong cation exchange (SCX). Generally one combines orthogonal chromatographic techniques to achieve maximum resolution of an analyte, such as SCX (charge separation) and reverse phased (hydrophobic separation) techniques. Though the combination of these methods has been successful, it has also exhibited limitations. As odd as it may seem that two reverse phased techniques (offline HPRP and online reverse phased (low pH) LC-MS) are combined in succession, the goal of the analysis, and the exact nature of the techniques in question reveal why such a combination is warranted.

At low pH, as is standard for online reverse phased separation (LPRP) coupled to mass spectrometry (LC-MS), histidine, lysine and arginine, glutamic acid and aspartic acid residues will be protonated. Thus peptides will have a net positive charge. In contrast at high pH (pH 10), the majority of the residues will be deprotonated, with the exception of arginine. In this manner, the exact peptide sequence affects its hydrophobic character, which contributes to the orthogonality of HPRP and LPRP. In addition, the goal of pre-fractionation is to reduce sample complexity across the LC-MS gradient, not necessarily to purify a particular set of peptides out the group of total peptides. In effect, the most successful techniques would be those that allow consistent peptide elution across the entire LC-MS gradient, while maintaining a reduction in sample complexity. Such fractionation would maximize the number of unique peptides isolated and fragmented during the MS duty cycle. With the sample pooling strategy outline in Figure 4.7, where early, middle and late fractions are combined from the high HPRP separation for LC-MS, this desired separation is obtained. A brief comparison of SCX and HPRP is discussed.

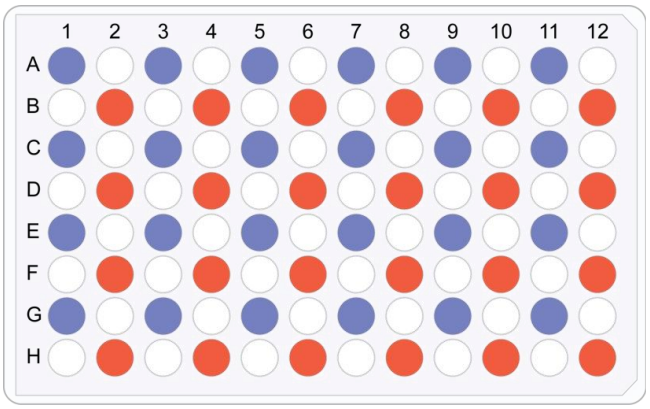


Figure 4.7. Diagram of HPRP fraction collection and pooling of fractions for MS analysis. 96 fractions are collected over a ~60 minute gradient (from ~10-70 minutes in the HPRP gradient, 38 second fractions) in a 96 well micro plate (1 mL wells). Twelve pooled fractions are created for LC-MS analysis by combining rows A, C, E and G for odd numbered samples (blue wells) and B, D, F and H for even numbered samples (red wells). This method of combining fractions limits the overlap between adjacent samples, increasing the number of uniquely identified peptides and proteins by LC-MS/MS. In contrast, a typical experiment relying on SCX-based pre-fraction utilizes 20-25 fractions, increasing LC-MS analysis time.

High pH reverse-phase (HPRP) columns separate peptides orthogonally, based on the hydrophobic character of a peptide and the frequency of basic residues in that peptide. In contrast, strong cation exchange (SCX) separates peptides primarily on positive charge. Though both SCX and HPRP separate peptides over roughly a one hour gradient in an effective manner (Figure 4.8, A), the HPRP methods tends to show more uniformity (based on the UV absorbance). In addition, HPRP tends to have better peak resolution; this resolution is likely due to faster partitioning of peptides between the mobile and stationary phases of the reverse phased surface vs. charge exchange surface of SCX. Typically, peptides identified by mass spectrometry have 2+, 3+ and to a lesser extent 4+ charge states, depending on which enzyme is used for digestion. These charge states tend to elute as group by SCX, and thus are not sufficiently resolved. Furthermore, peptide length and charge are correlated, and longer, highly charged peptides are not amenable to LC-MS/MS, rendering late SCX fraction less useful. These limitations are overcome by HPRP. The combined HPRP fractions (Figure 4.7) display a wider range of hydrophobicities compared to SCX during online LC-MS fractionation (Figure 4.8, B). Although later HPRP fraction will contain longer peptides as well, they will not necessarily be highly charged, rendering them amenable to mass spectrometry. The result of these properties is our ability to collect a greater amount of data by LC-MS using fewer fractions.

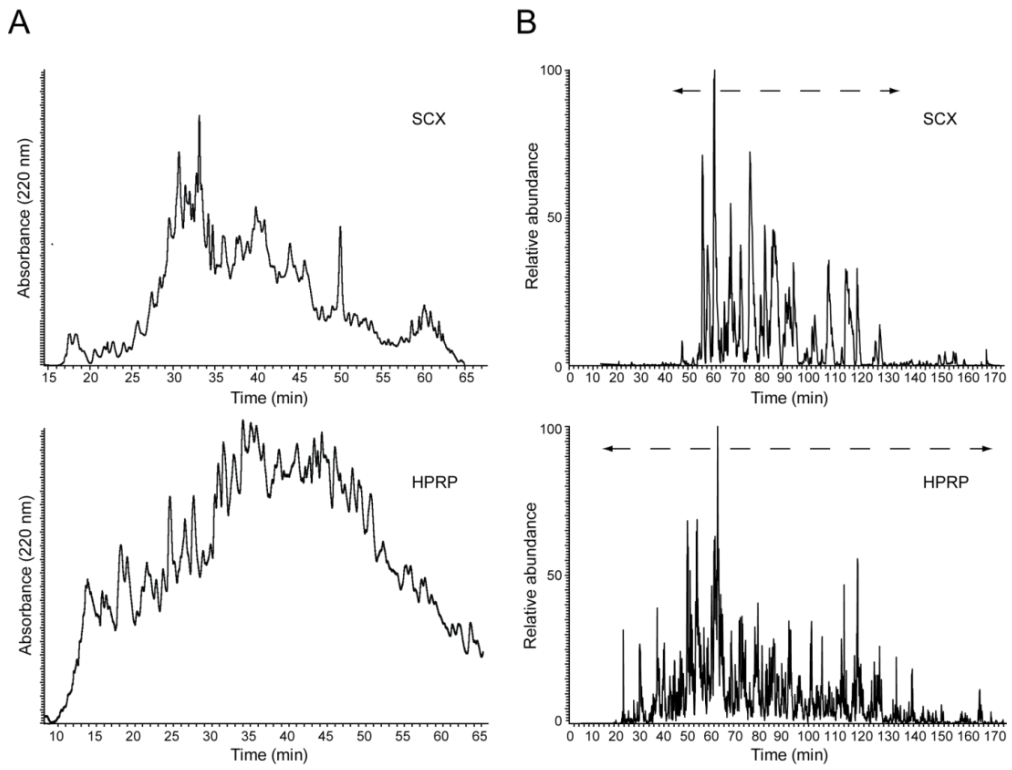


Figure 4.8. Comparison of peptides separated by SCX and HPRP. Both offline pre-fractionation (A) and online LC-MS (B) chromatograms are presented for SCX and HPRP samples. Both SCX and HPRP separate peptides over roughly a one hour gradient (A); HPRP, however, gives more uniform peptide elution over the full hour, whereas SCX show more pronounced tailing. HPRP also provides better resolution. HPRP experiments tend to be more reproducible in terms of peptide identifications as well. Compared to a typical SCX fraction, peptides from an HPRP fraction display a wide range of hydrophobicities throughout an LC-MS gradient (B, lower chromatogram). In contrast, peptides from an SCX fraction (B, upper chromatogram) elute over a smaller window. This "bunched" elution profile leads to fewer identification per LC-MS run. The use of HPRP leads to an increase in unique peptide and protein identification, and has become the preferred method for peptide pre-fractionation.

To demonstrate the discussed behavior on a proteome-wide scale, yeast proteomics experiments were performed using either HPRP or SCX. 12 pooled HPRP (discussed above) and 20 SCX fractions were collected and analyzed by LC-MS. One minute SCX fractions were used, and the total number was selected based on the elution profile. Total peptide, unique peptide and protein identifications were reproducible between HPRP fractions (Figure 4.9, A), whereas SCX fractions were highly variable (Figure 4.9, B). The HRPR method on average significantly outperformed the SCX method;

particularly with regard to unique peptide/total peptide ratio (77 +/- 2% vs. 47 +/- 11.3%) and with protein identification (average 1906 +/- 85 vs. 1062 +/- 470), HPRP is the preferred method.

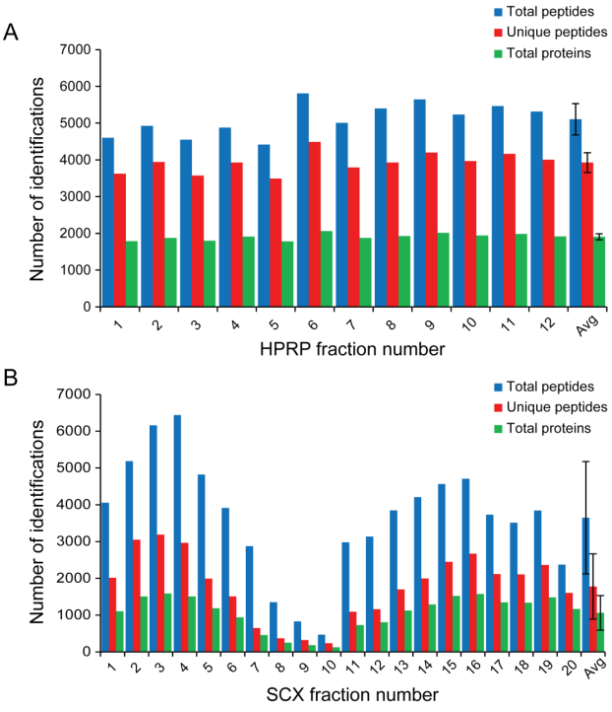


Figure 4.9. Comparison of identified yeast peptides and proteins between HPRP (A) and SCX (B). The "Avg" column displays the mean values of each category +/- one standard deviation. In this example, 12 HPRP and 20 SCX fractions were collected, encompassing the full gradient of eluted peptides in each case. Although certain SCX fractions (2-4) contain more total peptides, the HPRP fractions consistently contain more total peptides, and of greater importance, more unique peptide and proteins than the SCX fractions. The HPRP fractions have a high unique peptide to total peptide ratio (>75%), as compared to SCX (<50%). On average, an HPRP fractions contains 5000 total peptides, 4000 unique peptides, and nearly 2000 proteins; SCX fractions on average contain <4000 total peptides, <2000 unique peptides, and ~1000 proteins. The HPRP method also yields more consistency in peptide and protein identification (low standard deviation) between fractions, compared to SCX.

Improvements in TMT Reporter Ion Isolation and Quantification.

Though the introduction of the MS^3 method for TMT quantification was a technological achievement, it contained a major caveat; the gain in ion purity (removal of interference seen with MS^2 quantification, Figure 4.10, top) was overshadowed by a large reduction in reporter ion signal. The result of the signal loss is often poor quantification, where ions in a particular channel may be undetectable (Figure 4.10, middle). Thus when sufficient ions are present, an accurate ratio may be obtained; often, however, one or more channels are not present, which leads to quantification errors. The solution to the TMT signal problem was to coisolate multiple MS^2 fragment ions, in order to obtain a composite MS^3 spectrum with greater signal (Figure 4.10, bottom). This "multinotch" method on average increases TMT reporter ion signal by 8 fold (Figure 4.11, left), without typically introducing interference back into the

quantification (Figure 4.11, right)[13]. When comparing the MS^2, MS^3 and multinotch methods, we find that the multinotch consistently outperforms the two alternatives.

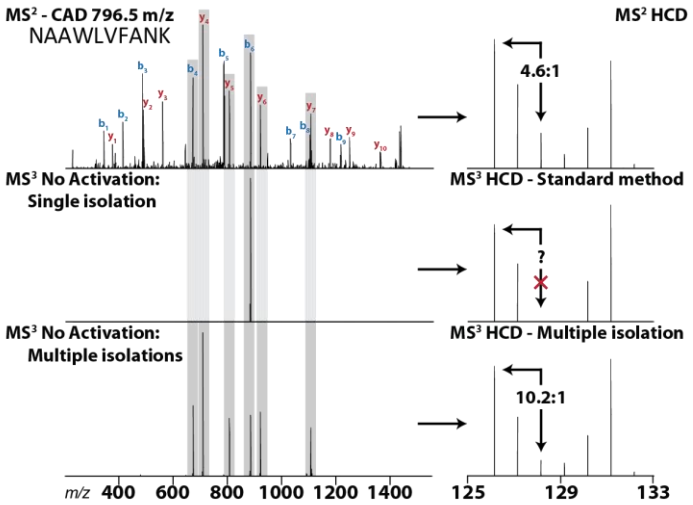


Figure 4.10. A multinotch method for MS^3 quantification avoids interference while increasing TMT reporter ion signal. As discussed the MS^2 method for TMT reporter ion quantification introduces interference, which compresses observed rations amongst the TMT channels (top panel, a 10:1 ratio is observed at 4.6:1). The MS^3 method abrogates interference, but does so at the price of lost TMT ion signal as the result of isolating a single MS^2 ion for HCD fragmentation (middle panel, missing value/undefined for the 10:1 ratio). The multinotch method employs a data-dependent algorithm to isolate multiple MS^2 ions for simultaneous HCD fragmentation. This method increases the TMT ion signal without increasing interference (bottom panel, the 10:1 ratio is observed properly). Adapted from *Isolating multiple MS^2 fragments using waveforms with multiple frequency notches improves MS^3 sensitivity ~8 fold over standard MS3-based TMT methods*, used with permission[13].

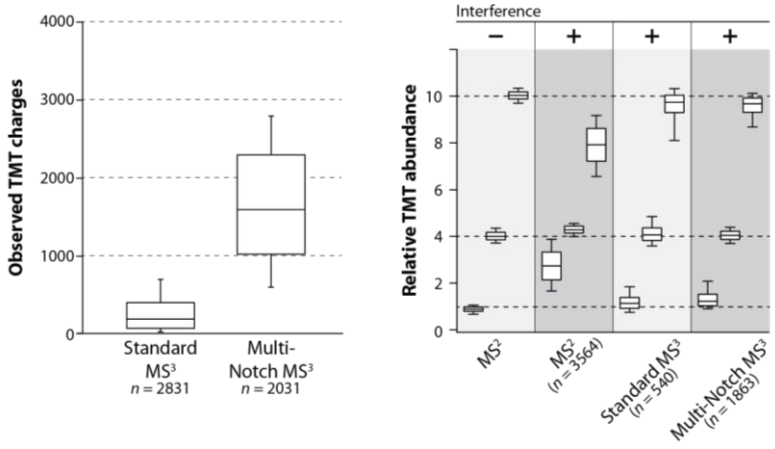


Figure 4.11. The multinotch method increases TMT reporter ion signal by an average of 8 fold which increase the accuracy of reporter ion ratios. The multinotch method increases the number of observed TMT charges in an MS^3 spectrum by an average of 8 fold, compared to the standard MS^3 method (standard box plots, left panel). The multinotch method does not typically introduce interference as observed in the MS^2 method (right panel), comparable to the standard MS3 method using a yeast/human model system (discussed by Ting et al[10]). Adapted from *Isolating multiple MS2 fragments using waveforms with multiple frequency notches improves MS3 sensitivity ~8 fold over standard MS3-based TMT method,* used with permission[13].

To assess the quantitative nature of the multinotch improvements, three yeast proteomics data sets of heat stressed vs. unstressed controls were obtained in biological triplicate (Figure 4.5). These data sets were obtained using the MS^2, MS^3 and multinotch methods of quantification. As the data were obtained in triplicate, several metrics of reproducibility can be analyzed: variance in stress/control ratios for each protein (TMT channels 129/126, 130/127, and 131/128), the channel signal variance (within a set of replicates, control or stressed, equal results obtain for either), or the peptide to peptide ratio variance (sum of the heat stress channels/sum control channels for a peptide). Each value is expressed as a percent of the coefficient of variance (CV, standard deviation divided by the mean, multiplied by 100). Using the protein ratio or channel signal metrics (Figure 4.12, A and B, respectively), the multinotch outperforms both the MS^2 and MS^3 methods, often displaying <5% CV. Intriguingly, the MS^2 method outperforms the MS^3 method in these cases; it is likely that interference is forcing each channel towards the same value (ratios of 1:1) is the MS^2 method, reducing the variance. Additionally the loss of signal in the MS^3 method contributes to its increased variance.

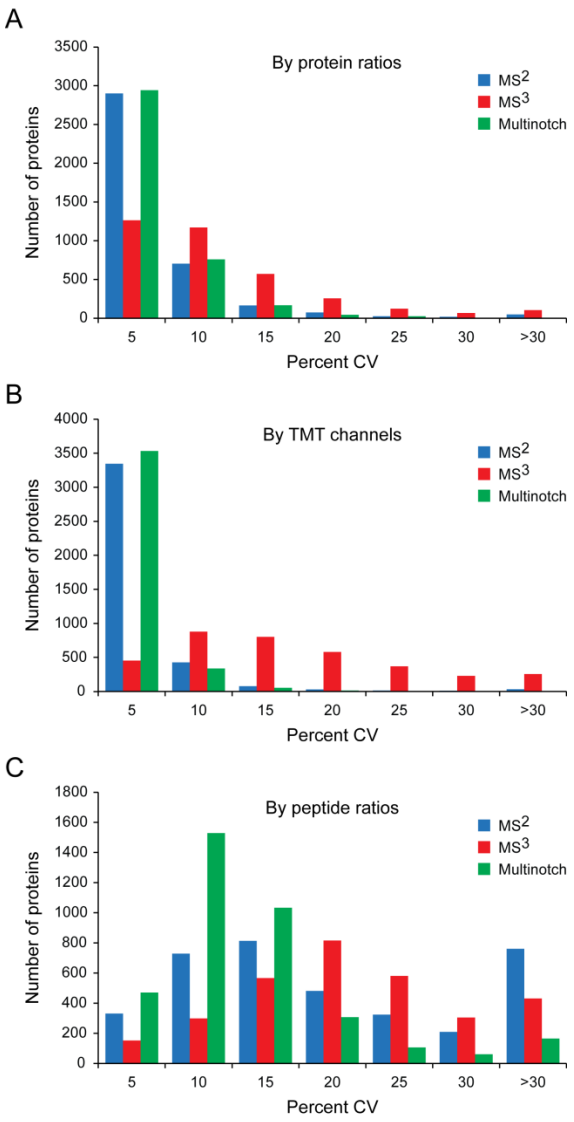


Figure 4.12. Demonstration of TMT ratio reproducibility amongst the MS^2, MS^3 and multinotch methods using the biological triplicate analysis of heat stress. 126-128 are the control channels and 129-131 are the heat stress channels. In all cases, histograms of the coefficient of variance (CV, standard deviation/average ratio * 100) are plotted, where a smaller percent signifies greater ratio stability among replicates. (A) Histograms using the 129/126, 130/127, and 131/128 ratios (at the protein level, using the summed channel intensity for all peptides from that protein). The MS^2 and multinotch method show more consistency amongst simple stress/control ratio, compared to the MS^3 method (most <5%). (B) Histograms using 126, 127 and 128 S/N measurements as a metric (protein level using the summed channel intensity for all peptides from that protein, peptide level analysis was identical in trend). Results were identical for the same plot using the 129,130 and 131 S/N. The MS^2 and multinotch method show more consistency amongst the stress and control channel S/N, compared to the MS^3 method (most <5%). In these measurements (A and B), the success of the MS^2 method is actually due to interference, as the compression effect artificially stabilizes channel to channel variation. The success of the multinotch method, however, is a result of the increased signal to noise measurements and true channel to channel reproducibility. (C) Histograms using the sum (129:131 S/N, heat stress)/sum (126:128 S/N, controls) for all peptides from a protein as a metric (peptide to peptide ratio reproducibility)**.** The multinotch method outperformed both the MS^2 and MS^3 method as a result of more accurate measurements of the true peptide TMT ratios. In many cases the MS^2 method had very high (>30%) CV, as a result of interference-based ratio distortion.

For the same reason, the MS^2 method performs the worst when comparing peptide to peptide variance (Figure 4.12, C). Though interference is common, many peptides are free of interference (or more likely interference only contributes a small amount of signal) in the MS^2 method; therefore, some peptides are quantified at near true values, whereas others are quantified at compressed values, leading to large variance. Using this same metric, the multinotch method outperforms the other two, and exhibits generally <10% peptide CV. The low variance observed with the multinotch is dependent upon signal to noise filtering to remove poor quantifications. A peptide filter, requiring that the sum of all

TMT channel (126-131) S/N equal at least 100 (empirically derived, see Figure 4.13 for details) greatly reduces peptide to peptide ratio variance. Such filters do not enhance MS^2 data (as interference is independent of TMT signal to noise) and are not feasible with non-multinotch MS^3 data (discussed below). Variance is reduced further in the multinotch method through the use of weighted averages (based on S/N, Figure 4.13), as is standard in our analysis.

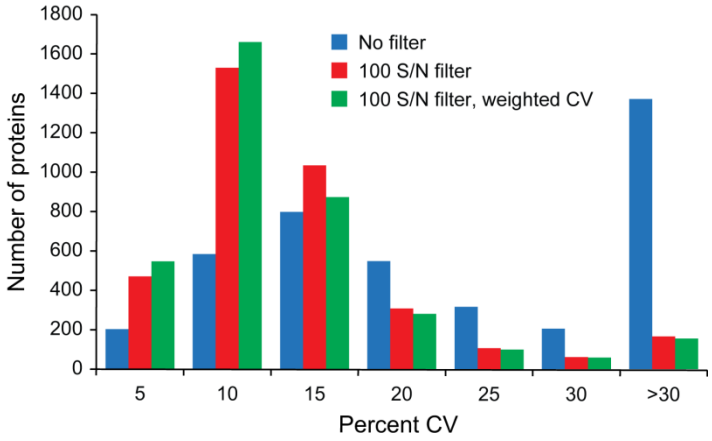


Figure 4.13. The use of signal to noise filters and weighted TMT ratios (weighted by peptide summed TMT S/N across all channels) decreases peptide to peptide ratio variance. Histograms of the CV using the sum (129:131 S/N, heat stress)/sum (126:128 S/N, controls) for all peptides from a protein as a metric are presented. A 100 S/N cutoff for summed TMT reporter ion intensity for a peptide (across all 6 channels, roughly 500 TMT charges) had previously been demonstrated to coincide with ratio stability. The 100 S/N filter significantly stabilizes peptide ratios for a given protein, compared to unfiltered data (peptide ratio CV is reduced from >30% to 10% or less in most cases). These data were obtained using the multinotch method; such filtering is not possible using the MS^3 method, as too many peptides fall below the cutoff. The use of weighted ratios based upon the summed TMT S/N of a peptide further reduces peptide to peptide ratio variance, though to a lesser extent, compared to the 100 S/N cut-off. Weighted CV is calculated from the weighted average and weighted standard deviation

In addition to reproducibility metrics, the multinotch method outperforms the MS^2 and MS^3 methods in terms of the relevant biological data. The multinotch achieves better statistics than the MS^2 and MS^3 methods (T-tests comparing stressed and unstressed channels, Figure 4.14, A). The relevance of statistics in proteomic analysis is elaborated upon later. As discussed above, the MS^2 method actually surpasses the performance of the MS^3 method in these tests due to compression in the MS^2 method, and poor S/N measurements in the MS^3 method. However, the MS^3 data set contains more biologically relevant proteins (traditionally considered relevant in high throughput studies, those that change by the

given fold change in Figure 4.14, B) compared to the MS^2 method. As expected the multinotch further increases relevant protein quantification over the MS^3 method.

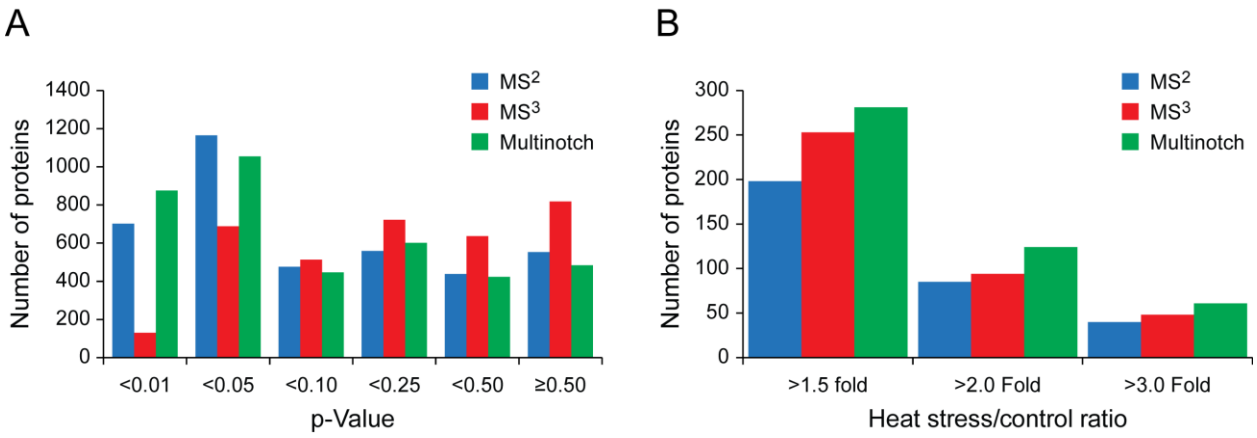


Figure 4.14. The multinotch method outperforms the MS^2 and MS^3 methods in identifying relevant biologically regulated proteins in heat stress. (A)Two-tailed T-test (126:128 vs. 129:131 arrays of summed peptide S/N from a protein) histograms after Benjamini–Hochberg correction (for multiple hypothesis testing) are plotted. The MS^2 method outperforms the MS^3 method, and the multinotch outperforms both the MS^2 and MS^3 method with respect to p-value. As with channel CV (Figure 4.12, B), interference artificially stabilizes channel to channel variance, thereby improving T-test performance. The exemplary performance of the multinotch method with respect to the T-test is true behavior and is as a result of increased S/N measurements. (B) Despite performing poorer with T-test, the data set using the MS^3 method contains a greater number of proteins whose heat stress to control ratio changes by a biologically meaningful value, compared to the MS^2 data set. As with the T-test, the multinotch method outperformed both methods with respect to protein ratios.

Combining Novel Proteomics Methods Offers a Significant Advantage Over Previous Standards

To demonstrate the advantages of combining the discussed novel techniques, two yeast proteomics data sets were compared; one was obtained using the SCX and MS^3 methods (the "standard" method), whereas the second was obtained using the HPRP and multinotch methods (the "novel" method). Notwithstanding any TMT S/N filtering, the novel method produces a richer data set compared to the standard method (Table 4.1). It contains more unique peptides and identified proteins, despite containing fewer total peptides (due only to the smaller number of fractions collected). Additionally, the novel method more faithfully quantifies the identified proteins, obtaining quantitative data for >95% of the identified proteins.

Table 4.1. Data set statistics from a yeast stress experiment performed using SCX and single notch MS3 vs. HPRP and multinotch MS3, no data filtering implemented. Each data set has been filtered to a 1 % protein level FDR. Peptides were collapsed into the minimum number of proteins required to explain all observed sequences, and those peptides with signal in at least one TMT channel were considered quantified. All peptide quantifications are considered unique quantification events with MS^2 and MS^3 based quantification, and were therefore used in the calculation of protein quantification. No TMT reporter ion filters have been applied to these data sets. In all cases the HPRP/multinotch method outperformed the SCX/MS^3, with the exception of total peptide identifications; though the ratio of total peptides to fractions analyzed is still higher for the HPRP/multinotch method.

Experimental conditions	Total peptides	Unique peptides	Total proteins	Quantified peptides	Quantified proteins
SCX, 20 fractions, single notch MS3	72, 903	26, 174	3, 724	37, 977	3, 348
HPRP, 12 fractions, multinotch MS3	61, 173	33, 654	4, 005	48, 953	3, 831

When the aforementioned signal to noise filters were applied, the standard method becomes virtually unusable (Table 4.2). Only ~30% of the previously quantified peptides, encompassing ~75% of the previously quantified proteins remain. Many of those proteins that remain are quantified using few peptides. In contrast, ~80% of the previously quantified peptides, encompassing >95% of the previously quantified proteins remain in the data set obtained using the novel method. In the novel method, a protein is quantified using more than ten peptides on average, whereas in the standard method, this number is less than five. Though there still is a small reduction in the number of quantified peptides and proteins in the novel data set upon the implementation of signal to noise filters, the quality improvements achieved are paramount. It is clear the novel method outperforms the standard method in all facets of the analysis. Of importance, the novel method data set was acquired using only 12 HPRP fractions, compared to the 20 SCX fractions in the standard method data set. Therefore, the discussed gains in data set quality are on top of a 40 % reduction in analysis time. These data impress upon the need for technological improvements, and the advantages such improvements offer.

Table 4.2. Data set statistics from a yeast stress experiment performed using SCX and single notch MS3 vs. HPRP and multinotch MS3, implementing S/N filters. Each data set has been filtered to a 1 % protein level FDR. As in table 1, peptides were collapsed into the minimum number of proteins required to explain all observed sequences. Peptides were required to have a summed (across all 6 TMT channels) reporter ion signal to noise ratio >100. This cutoff value had been observed to remove the variability associated with poor quality quantification events. Those peptides passing the S/N cutoff, with signal in at least one TMT channel were considered quantified. All peptide quantifications are considered unique quantification events with MS^2 and MS^3 based quantification, and were therefore used in the calculation of protein quantification. Using the multinotch method, only a small fraction of the data is thrown out (20% of the peptides and 4% of the proteins removed). Conversely, without the multinotch method, 70% of the peptides and 25% of proteins are thrown out, due to the low TMT signal to noise. The application of the multinotch method is required for robust quantification using TMT and is now the preferred quantification method of TMT labeled peptides.

Experimental conditions	Total peptides	Unique peptides	Total proteins	Quantified peptides	Quantified proteins
SCX, 20 fractions, single notch MS3	72, 903	26, 174	3, 724	11, 988	2, 494
HPRP, 12 fractions, multinotch MS3	61, 173	33, 654	4, 005	38, 350	3, 672

Applying Novel Proteomic Methods to the Multiplexed Analysis of the Yeast Stress Response

As detailed in Figure 4.5, three yeast stress data sets were obtained using the methodological improvements discussed. These data sets included a biological triplicate analysis of heat stress, a heat stress time course, and a comparison of five yeast stress states to an unstressed control. These data sets demonstrate the three main types of proteomics experiments which are made possible through quantitative multiplexing: statistical comparisons, temporal analyses and discovery-based experiments. The use of these data sets is twofold: first, due to the proteome wide nature (large proteome coverage) and quality of quantification, these data are useful a for systems level analysis of the yeast stress response. The wealth of data contained in these experiments is important for the generation of new hypotheses. Secondly, the data is useful as a community resource on the protein level adaptation to environmental stress. The great quantity of yeast stress data available typically has focused on genetic responses to environmental stress[23]; often there is a disconnect, due to post transcriptional regulatory mechanisms, between transcript and protein abundance[24]. Accordingly, these data provide a useful tool

for the direct analysis of the stress response, and is the largest yeast proteomic analysis implementing quantitative multiplexing technology.

Stress Data Set Statistics and Comparisons

In each stress experiment ~30, 000 or more unique peptides were identified, corresponding often to >4000 proteins (Figure 4.15, A unique peptide IDs, B proteins IDs). After all data quality filters were implemented, typically ~3, 700 – 4, 000 proteins were quantified in each experiment (Figure 4.15, B). The vast majority of proteins from all experiments were both identified and quantified using multiple peptides, often with >10 (figurer 15, A identifications, B, quantifications).

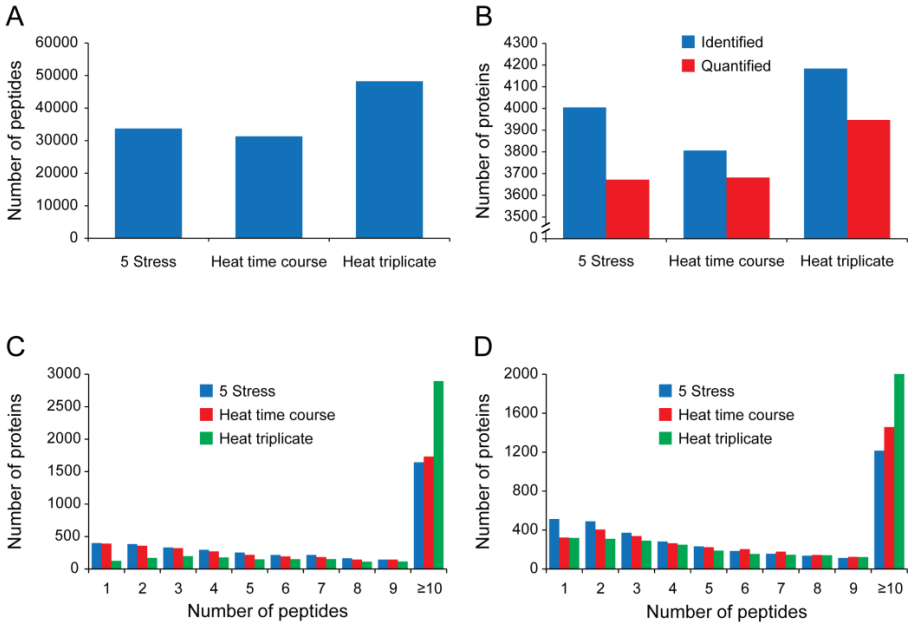
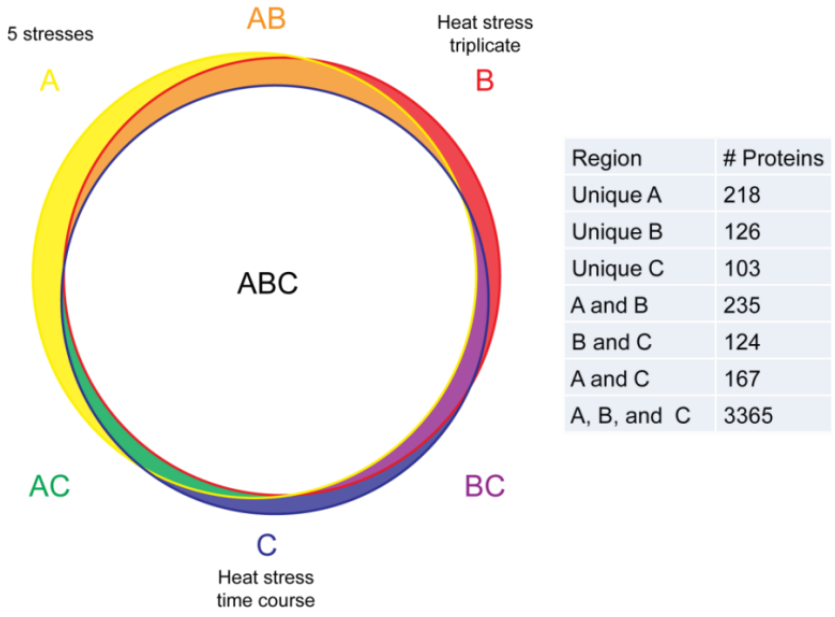


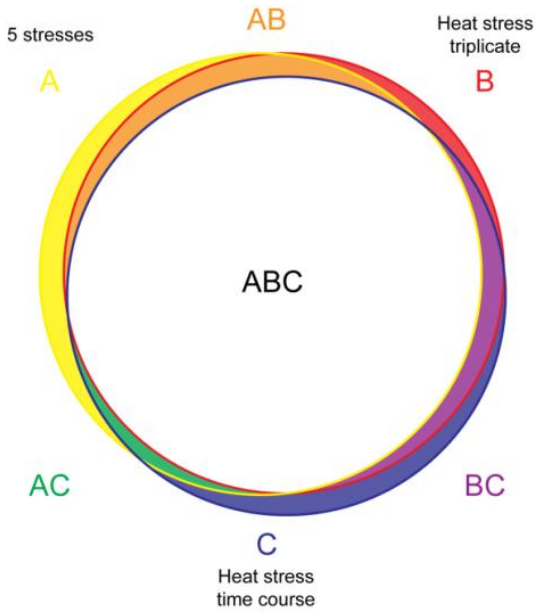
Figure 4.15. Data set statistics from the three yeast stress experiments. (A) In all experiments, nearly 30, 000 or more unique peptides were identified, providing a large amount of protein coverage. (B) In all experiments nearly 4,000 or greater protein were identified, demonstrating the proteome wide character of these analyses. The vast majority of identified proteins (nearly 95%) were quantified in al analyses. Of the proteins identified in each experiment, 90 % or greater were identified with (C) and quantified with (D) multiple peptides. Many of the proteins identified and quantified were done so with >10 peptides (C, D) demonstrating the depth of this analysis and by extension its accuracy (greater numbers of peptides/protein correlates with stable quantification ratios). The biological triplicate analysis of heat stress contains technical replicate data (additional LC-MS analyses of the same 12 HPRP fractions) out of necessity; during the first set of replicates the instrument was underperforming, which led to poorer TMT S/N. Many peptides were filtered out with the 100 S/N cutoff. The fractions were re-shot on a separate LTQ-Orbitrap Velos, also with the multinotch algorithm implemented. The initial data, however, was included, as the peptides that did pass the cutoff were still of high quality. Hence, the biological triplicate analysis of heat stress contains greater proteome coverage, as observed in this figure.

In addition to high quality data within a given experiment, there was a large overlap between identified (Figure 4.16) and quantified (Figure 4.17) proteins between experiments. In cases where technical replicates had been obtained (triplicate heat stress, explained in Figure 4.15), only 12 of the total LC-MS runs were used in overlap comparisons, so that equal comparison could be made. Generally there was a ~85% overlap in protein identification between any two experiments. There was an 80% overlap between all three experiments, comprising 3, 365 proteins. This number of overlapping proteins between three experiments is nearly twice that observed in comparable a TMT experiments[25]. In total 4, 235 proteins were identified (separately obtained to control FDR), representing one of the largest single yeast proteomics data sets to date.



Region	# Proteins
Unique A	218
Unique B	126
Unique C	103
A and B	235
B and C	124
A and C	167
A, B, and C	3365

Figure 4.16. Overlap in protein identification between experiments. In each data set, the list of identified proteins was collapsed into the minimum number of proteins required to explain all peptides. Redundant (by sequence) peptides were assigned to the most likely protein, the protein with the larger number of spectral counts. The equivalent of 18 protein identification experiments is presented here. The yellow circle (A) represents proteins identified in the five stress experiment; the red circle (B) represents proteins identified in the biological triplicate heat stress experiment; the blue circle (C) presents proteins identified in the heat stress time course experiment. In all cases only 12 HPRP fractions were used for each experiment to ensure equal comparisons. The large overlap in protein identifications between experiments allows for unprecedented biological comparison between a great number of samples. Without TMT and HPRP separation, this level of overlap would not be possible while collecting such a comprehensive body of quantitative data. 4232 proteins were identified in total (1% protein level FDR), representation the largest analysis of its kind.



Region	# Proteins
Unique A	180
Unique B	162
Unique C	141
A and B	214
B and C	153
A and C	218
A, B, and C	3125

Figure 4.17. Overlap in quantified proteins between experiments. Proteins were collapsed and redundant peptides were assigned as previously stated. Peptides were required to have a summed TMT reporter ion signal to noise of 100 across all 6 channels to be considered quantified to ensure high quality quantification. The equivalent of 15 binary comparisons is represented here. The yellow circle (A) represents proteins quantified in the 5 stress experiment; the red circle (B) represents proteins quantified in the biological triplicate heat stress experiment; the blue circle (C) presents proteins quantified in the heat stress time course experiment. As previously stated, only 12 HPRP fractions were used for each experiment to ensure equal comparisons. The equivalent of over 60, 000 (over 50, 000 stress/control ratios) protein measurements (and hundreds of thousands of peptide measurements) is contained in the three data sets. These experiments encompass the largest quantitative data set of its kind in yeast, where a total of 4, 178 proteins were quantified. In contrast to other published TMT data sets, this data set applies rigorous quantitative data filtering (based on S/N) and is free of interference.

More relevant than protein identification overlap, however, is the overlap between the quantified data. Since all the experiments are related through their analysis of yeast stress, shared quantified proteins between experiments permit both technical (e.g. ratio variability) and biological comparison. Generally there was an 80% overlap in quantified proteins between any two experiments. There was a 75% overlap between all three experiments, comprising 3, 125 proteins, over twice that observed in comparable TMT experiments[25]. In total 4, 178 proteins were quantified between all 36 samples (4, 357 using all data generated for this chapter). These proteins were quantified in the absence of interference and with rigid reporter ion filters; few if any TMT data sets exist with such quality

control. The equivalent of over 60, 000 (over 50, 000 stress/control ratios) protein measurements is contained in the three data sets, demonstrating the tremendous effort a comparable data set would require by other means. Such efforts would not be feasible SILAC, for example, an alternative LC-MS based quantification technique. Using the discussed level of overlap between two experiments for protein identification and quantification (0.85 and 0.80, respectively), the 15 required SILAC experiments would contain a final overlap of 0.85^{14} = ~10% and 0.80^{14} = ~4%. Realistically, however, a core group of ~1500-2000 proteins which are consistently found in a typical yeast experiment would be quantified. Additionally, many months vs. a week of time would be dedicated to such experiments.

The Use of Statistics Permits a Deeper Understanding of Protein Regulation

It is difficult to determine from a ratio alone how significant a particular protein may be in the heat stress response. Particularly at lower magnitude changes (<2 fold), assigning functional relevance to a protein would require prior knowledge of its role in the biology at hand. Such blind assignment requires knowledge of protein ratio variance to determine significance. Histograms of heat stress/control ratios show little obvious change based solely on magnitude (Figure 4.18, A), except for proteins which change by a large fraction (several fold). Indeed the vast majority of the proteome appears unchanged with heat stress. However it may be as important for a cell to modulate many proteins at small magnitudes simultaneously, as it is a few proteins at high magnitude.

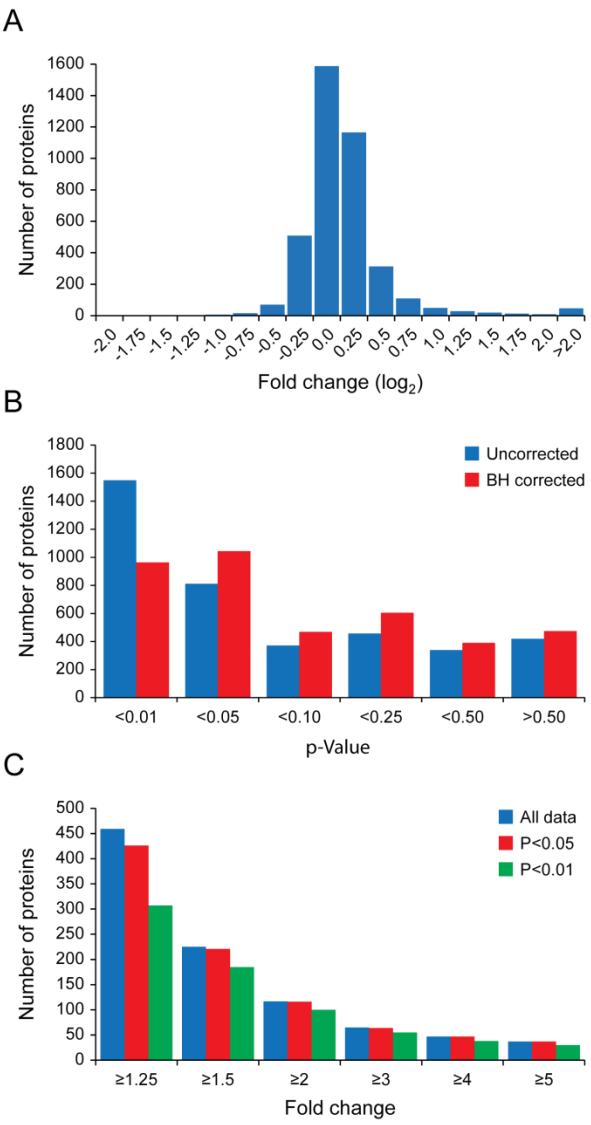


Figure 4.18. Data set statistics from the biological triplicate analysis of heat stress. (A) A histogram of the heat stress/control ratio (Sum (129-131)/Sum (126-128) TMT ions, Log_2) demonstrates that the vast majority of proteins (~90%) do not change upon heat stress by large ratios; additionally, a greater fraction of proteins which are regulated by heat stress are upregulated compared to down regulated. (B) Two-tailed T-test (126:128 vs. 129:131 arrays of summed peptide S/N from a protein) histograms after plotted either directly or after Benjamini–Hochberg correction (for multiple hypothesis testing). Although the Benjamini–Hochberg correction shifts p-values toward larger number, many proteins (nearly 1,000) still pass a T-test cutoff of 0.01. The accuracy of TMT quantification is responsible for this exemplary performance, as the variance between biological replicate is low. (C) The majority of proteins (only upregulated plotted) that change by 1.5 fold or greater pass corrected T-test cutoffs of 0.05 and 0.01. Many proteins that change by ratios as small as 25% also pass a stringent cutoff of 0.01. Thus many of these proteins that may be ignored in other analyses (such as a SILAC analysis) are statistically relevant and may collectively help describe the biology of heat stress resistance. Indeed cumulative or cooperative change, however small individually, may be relevant.

Traditionally, distribution statistics have been used as metric for assigning significance to proteins in a proteomic data set[26]. In these analyses, the burden of assessing reproducibility is often left for post experimental validation, such as by western blotting. The cutoffs are generally set at 2 standard deviations or greater, which although useful for identifying regulated proteins, are generally harsh and non-specific to the quality of quantification; moreover, much of the data set is removed with such simple filters, and the data that remains can often be obvious. In this analysis, a two standard deviation change was equal to two-fold; 128 proteins changed by a margin of two-fold or greater, though the

majority of proteins (117) were upregulated. This contrast between upregulated and downregulated proteins highlights an additional issue with simple cutoffs, that the biology at hand or the presence of tailing may be ignored when using distribution statistics alone. It is evident in the heat stress response, that upregulated heat stress protectant proteins are required at much greater magnitudes than those downregulated proteins which may be deleterious in the heat stress response. As is discussed, however, there is a large group of biologically related proteins (e.g. ribosome machinery), which are finely regulated (< 2 fold change); these proteins may be important in the heat stress response, most of which would be lost using distribution statistics alone, and may not be visible by techniques such as western blotting. In contrast to distribution statistics, the use of TMT for biological triplicate analysis adds a component of reproducibility, which is readily useful for performing more rigorous statistical tests (such as the T-test) between the treated (here heat stress) and control samples. Replicates can distinguish consistent biologically relevant changes from those changes due to experimental noise or poor quality quantification.

In this analysis, nearly 1000 proteins were extracted with significant p-values (<0.01, Figure 4.18 B, T-test p-values after Benjamini-Hochberg correction for multiple hypothesis testing[16]). Many proteins were discovered to change by a variety of magnitudes, from over 450 at a 1.25 fold or greater change, to over 50 at a 3 fold or greater change (Figure 4.18, C). Without statistics it would be particularly difficult to assign significance to these subtly changing proteins, and they would generally be disregarded. Conversely, if such a lenient cutoff of a 1.25 fold change were arbitrarily applied without statistics, a large false discovery rate of regulated proteins would be generated; roughly 1/3 of those proteins whose ratio changed by 1.25 or greater showed no statistical significance (Figure 4.18C, $p < 0.01$ vs. all data). Generally those proteins that changed by larger ratios (>2 fold) were statistically significant. The reason why many of these subtle changes were found to be significant lies at the inherent reproducibility of TMT quantification (once the discussed S/N filters have been applied, Figure 4.13).

The general reproducibility of quantification between biological replicates in both the control and heat stress channels (Figure 4.19) allows proteins with smaller average heat stress/control changes to pass a significance threshold of 0.01.

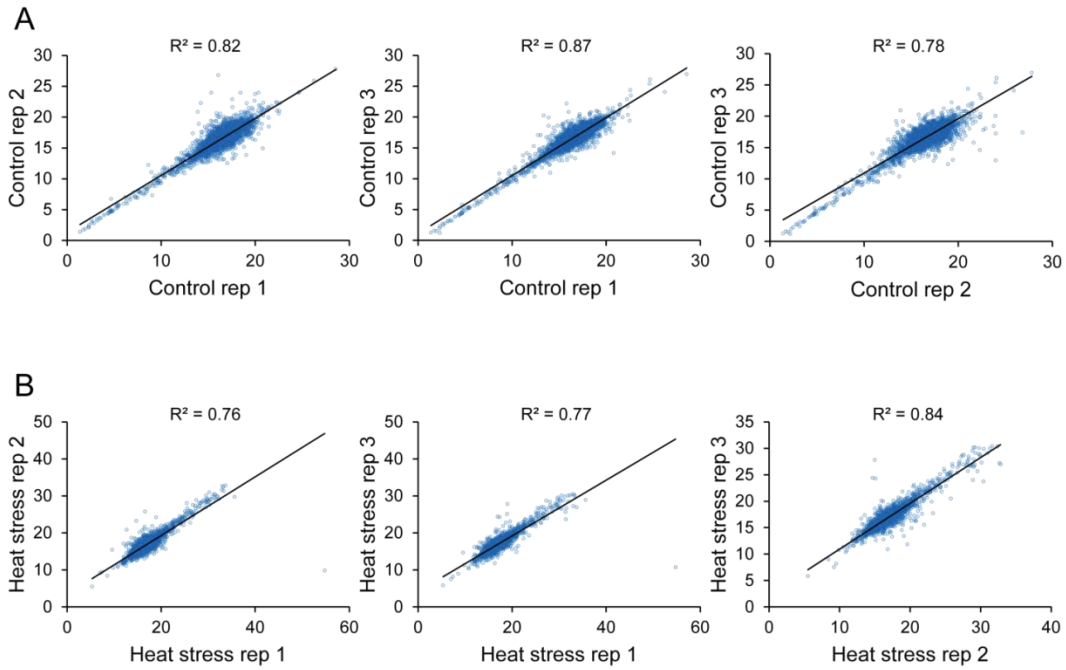


Figure 4.19. Relative TMT intensities are reproducible amongst the control samples (A) and heat stress samples (B). There is a high degree of reproducibility amongst the replicates, positively contributing to high quality statistical analysis presented. Only proteins which were quantified with > 1 peptide are plotted, which removes the vast majority of outliers (though few exist). Of importance, the one obvious outlier which remains (YLR392C, outlier in heat stress replicate 1 at a value of ~55) greatly affects the correlation between heat stress replicate 1 and the other heat stress replicates. When it is removed, the R^2 correlation improves to 0.85, demonstrating the deceptive nature a single outlier has on linear regression.

Filtering the data by p-value and plotting the relative S/N of control or heat stress channels against one another for all proteins (e.g. heat stress replicate 1 vs. heat stress replicate 2, Figure 4.20), respectively, demonstrates this correlation in greater detail. Even with a modest p-value cutoff of 0.1, two of the replicates are highly similar (Figure 4.20, B, $R^2 = 0.93$). With a more stringent p-value cutoff of 0.01, the two replicates are virtually identical (Figure 4.20, C, $R^2 = 0.98$).

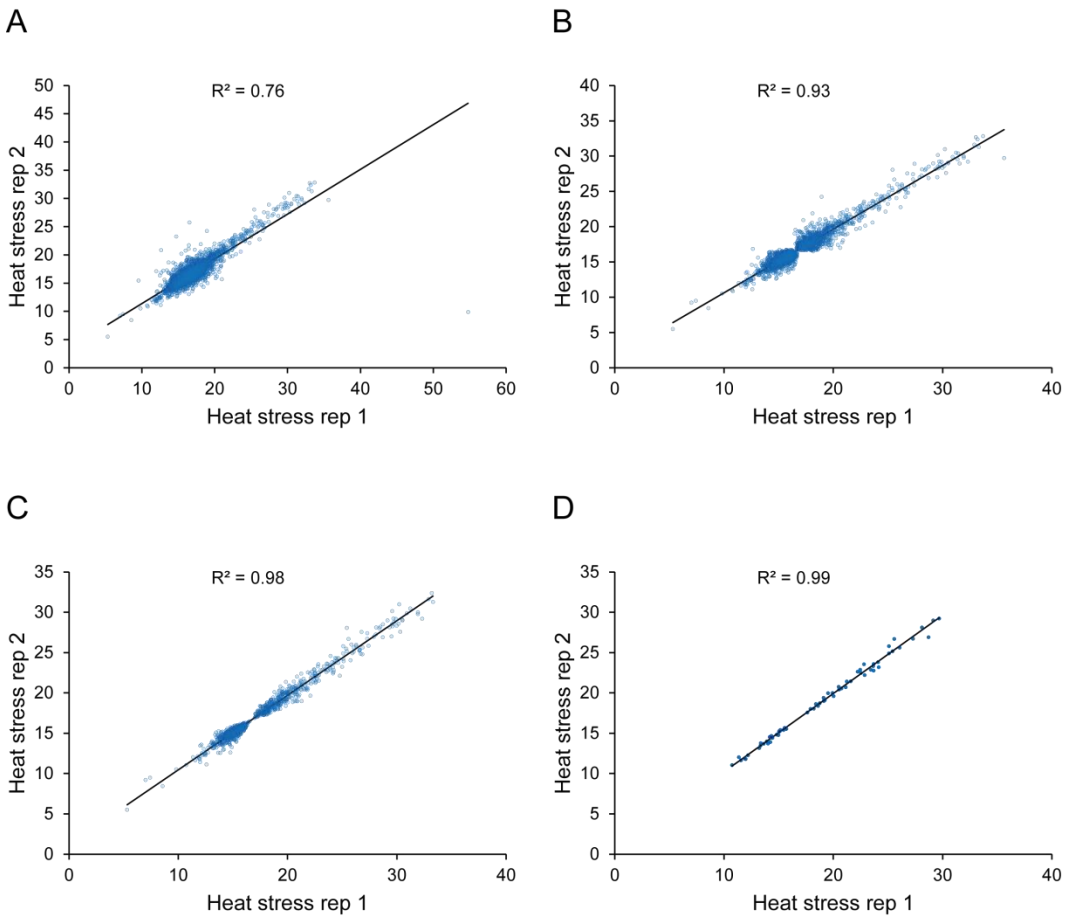


Figure 4.20. Statistics provide an effective means of removing variance between replicates. The comparison of heat stress replicate 1 and heat stress replicate 2 is presented. As was presented in Figure 4.19, these two replicates demonstrated the lowest degree of correlation (with no filter plotted in A), and therefore would benefit the most from additional filtering. (B) Using a p-value filter of 0.05, the correlation drastically improves to 0.93. With a p-value filter of 0.01 (C) or 0.001 (D), the samples become virtually identical.

Histograms of protein heat stress/control ratio for proteins passing p value cutoffs of 0.05, and 0.01 (Figure 4.21) demonstrate that both significant upregulated and significant downregulated proteins are present in the data. As the p-value filter is shifted lower (from 0.05 to 0.01 in this figure), two distributions begin to be observed (median values at ~-0.25 and 0.5 $\log_2$ ratios). Thus it is clear that the distribution of regulated data (those significant in the T-test) are not normally distributed, highlighting the reason why simple distribution statistics may fail in such an analysis. The distribution of all data appears normal, however, permiting the use of discussed statistical tests.

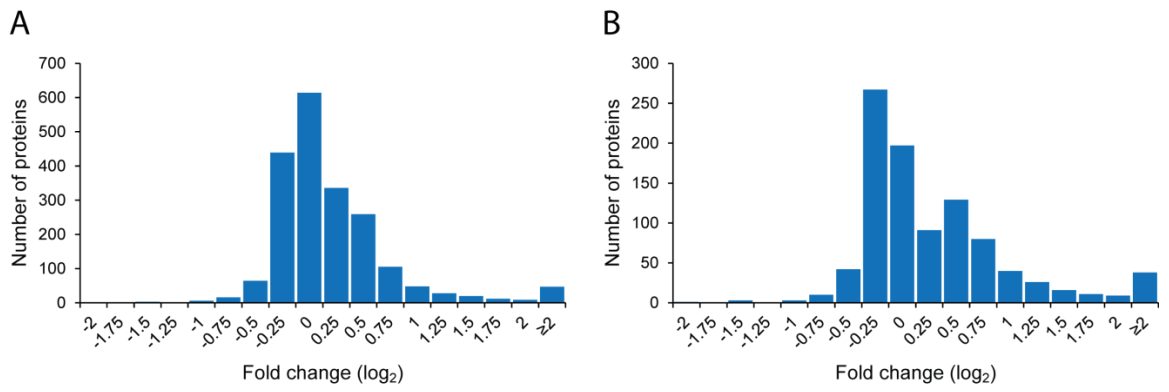


Figure 4.21. Distribution of protein ratio (log_2 values) for proteins passing a corrected T-test of 0.05 (A) and 0.01 (B). As in Figure 4.18, A, both using a corrected T-test cutoff of 0.05 (A) and 0.01 (B), more proteins were observed to be upregulated compared to downregulated. This observation is likely biologically significant and may indicate that upregulation may be an active process, whereas downregulation may be passive (perhaps simply based on protein turnover after transcription/translation is ceased). However, it is possible this behavior is a reflection of the normalization scheme, in that protein upregulation must be balanced by apparent downregulation so that all samples have the same total protein amounts; though possible, being that a majority of proteins do not change, this effect is unlikely. Proteins passing a p-value cutoff of 0.01 show a bimodal distribution, likely a result of the balance between channel reproducibility and heat stress/control ratio differences. Those proteins with large ratios are more likely to be statistically significant, as are the proteins with very reproducible signal amongst the replicates. Those proteins in the middle range, whose heat/stress ratio is not large and whose channel reproducibility is not impeccable will not pass a stringent T-test cutoff of 0.01.

Biological Interpretation of Heat Stress Proteomic Data

Due to the proteome wide nature of the analysis, the quality of the data, and the use of replicates, we have an unprecedented ability to analyze heat stress on a global scale. This analysis was undertaken without prior knowledge of the system of heat stress, yet many important components and pathways within the system, described for a number of years of smaller scales, were identified. Here I seek to discuss the protein expression trends that are primarily responsible for separating out heat stress from steady state controls, and to remark on the combination of protein level processes which may be responsible for the ability of yeast to adapt to high temperatures.

Expression Patterns Separate Heat Stress and Control States into Two Primary Components

Often the first step for interpreting large scale data involves a means of assessing similarities and differences between the biological states as a result of global protein expression patterns. Some of

these methods have been discussed and include data clustering (hierarchical, k-means, etc.) and principal component analysis (PCA). A dendrogram of the sample array reveals that the control and heat stress samples cluster as two primary groups, as would be expected based on the experimental design (Figure 4.22, A, distance is Euclidian). Of the replicates, heat stress replicates two and three, and control replicates one and three were most similar. Methods of dimensionality reduction such as PCA are particularly useful for multivariate data sets, such as those generated through mass spectrometry based proteomics, due to the high dimensionality of the data. Here, the data is separated into two primary principal components which explain 91% and 3 % of the total variance, respectively: component one separates the heat stress samples from the controls in all replicate equally, and component two explains some of the variance associated with biological stochasticity (although it also could represent consistent technical variation between samples). The biological variation which comprises component two is primarily a result of differences observed in control sample two and heat stress sample one, the two samples which clustered separately amongst the heat stress and control channels. The remaining components, which explain ~5% of the sample variance, show no clear trend amongst the replicate, and likely are a representation of noise in the data. This low amount of unexplained variance is indicative of a small level of technical stochasticity among replicates.

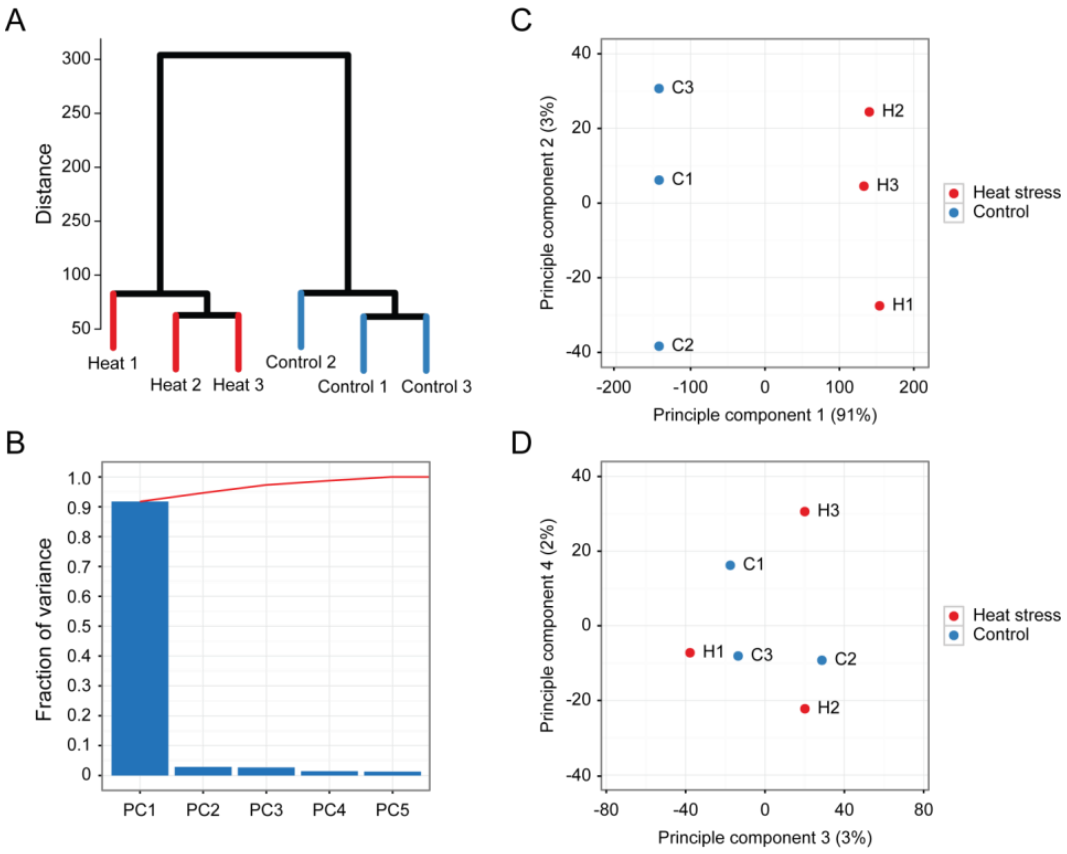


Figure 4.22. Heat stress vs. no treatment explains the majority of variance observed in the data set, by principal component analysis. (A) A hierarchical cluster tree of the sample arrays demonstrates a high level of similarity between replicates and a large difference between treatment groups. Each array was composed of the relative TMT S/N across all proteins (so that each protein array sums to 100% across the 6 samples), and clustering was carried out using the Euclidian distance metric and centroid linkage clustering method (though other methods were extremely comparable). (B) Principal component analysis revealed that the majority of variance was described by one component (principal component one, PC1, >90%). (C) PC1 represents heat stress vs. untreated yeast, and faithfully separates all control from all stressed samples. Principal component two (PC2), though responsible for only a small fraction of the variance in the data set (3%) may represent inherent biological variability, as the replicates are separated along component two independent of treatment condition. The reaming components likely reflect some level of experimental noise as no obvious trend is present, and represent the remaining ~5% of variance. Nearly no signal was present in component six, and it was therefore omitted.

Each component contains a matrix of loading values (one for each protein), related to the original values of that protein in each experimental condition. Plotting the loading values from one component against those of another component can be a useful means of identifying which proteins make large contributions to a given principal component. The usefulness of such plots, however, is dependent upon defining the biological or technical relevance of each component (previously discussed)

as well as the separation achieved on each axis for a given protein. A plot of the loading values for PC1 and PC2 (Figure 4.23) in this analysis reveals several points which separate from the large cloud of data centered at coordinates [0,0] in both positive and negative directions (though separation in positive direction is more apparent). Relevant proteins comprising different loading values along PC1 and PC2 are labeled to highlight proteins which make significant contributions to the differences observed in heat stress, and those potentially due to biological variation (see Figure 4.23 and 4.24 for details).

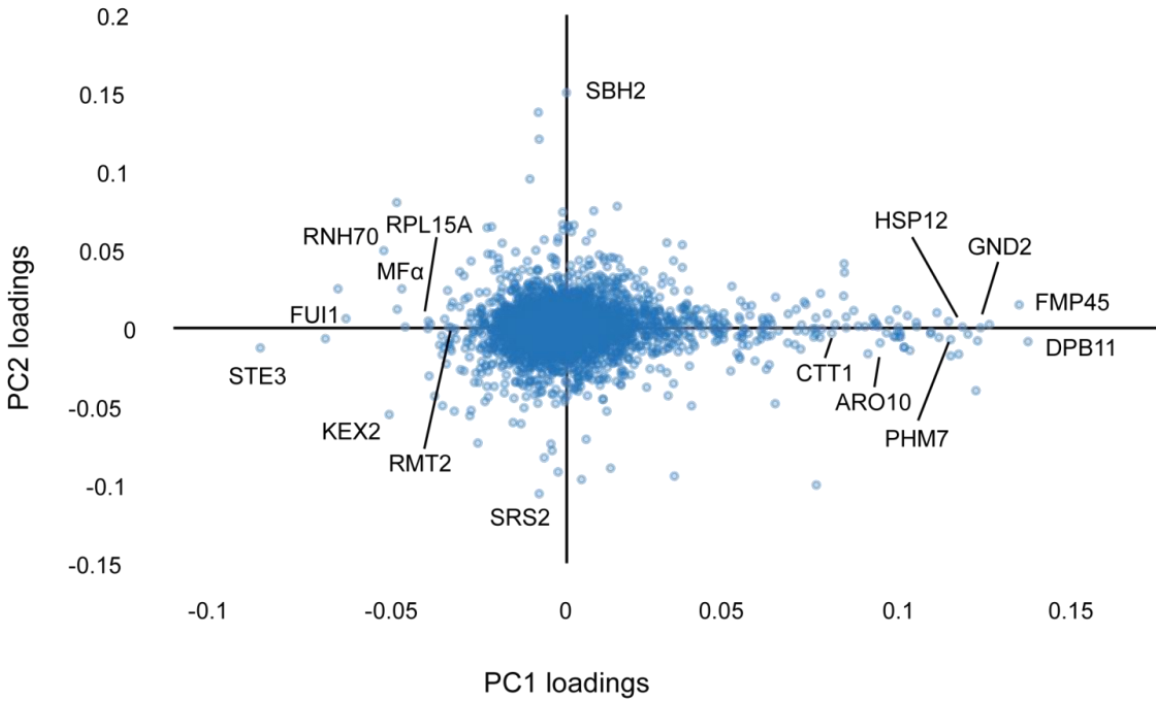


Figure 4.23. PCA loadings for PC1 plotted against PC2 separated upregulated and downregulated proteins. PC1 separated upregulated (positive direction) and downregulated (negative direction) proteins from one another. PC2 to some degree resolved the more consistent quantifications, though not in a statistically meaningful way (it does not seem to strongly correlate with p-value, data not shown), See Figure 4.22, C, for the distribution of the samples amongst the components, and Figure 4.24 for the specific normalized TMT ion distributions of the exemplified proteins. Highlighted hits encompassing different loading values are shown, many of which have large relative loading values, and are thus major factors in the principal component analysis. These highlighted hits encompass biological categories such as rRNA/ribosome related (downregulated), protein folding and metabolism (upregulated). One obvious outlier (at ~ [0.07, -0.1], YLR392C) is actually the same outlier observed in Figure 4.19, and is responsible for a significant portion of the variance observed between replicates, composing PC2.

The normalized TMT intensities for each are plotted (Figure 4.24) for these highlighted protein (Figure 4.23). Consistent quantification across the control and heat stress channels is generally observed for those highlighted in PC1. The positive component loading values correlate with protein upregulation upon heat stress (ratios from nearly exclusive heat stress to ~5:1, heat stress/control, are exemplified). The negative component values not surprisingly correlate with proteins which are downregulated upon heat stress, though the magnitudes of change are much reduced (ratio from ~5:1 to ~2:1 are highlighted). The highlighted, upregulated proteins are involved in such processes as protein folding, oxidation-reduction and metabolism. The downregulated proteins are generally involved in with the ribosome (proteins components and rRNA biogenesis) and mating. The proteins SBH2 and SRS2, highlighted as outliers amongst component two, are intended to exemplify biological variation (there are inconsistencies amongst the control and heat stress channels in these examples) and do not seem to be involved with the stress response. Though component two does seem to comprise some biological variation, generally the variation amongst replicate is low, limiting the usefulness of this component within the context of the whole data set. As noted, component two contain only 3% of the variance in the data set. Additionally, the aforementioned T-test filtering removes any highly variable protein quantifications from the final list of proteins deemed to be regulated by heat stress.

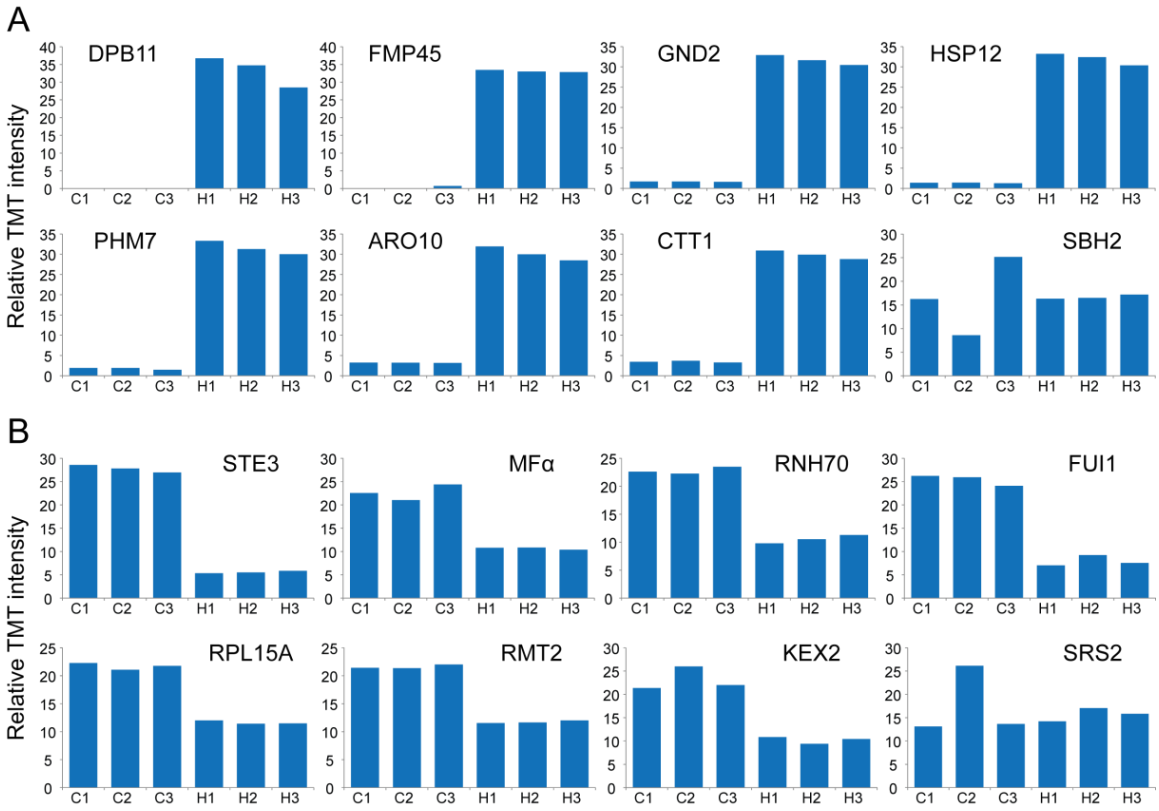


Figure 4.24. Normalized S/N of TMT ions of the highlighted proteins from Figure 4.23. (A) Proteins separated by PC1 in the positive direction are upregulated. All highlighted examples are reproducibly quantified across all six samples. DPB11 (DNA damage/replication, MEC1 activator), FMP45 (mitochondrial protein, involved in sphingolipid maintenance), GND2 (6-phosphogluconate dehydrogenase), HSP12 (small heat shock protein), PHM7 (phosphate metabolism), ARO10 (phenylpyruvate decarboxylase, amino acid catabolism), and CTT1 (catalase, oxidative damage resistance) exemplify major factors separating out heat stress from the control due to increased protein abundance. (B) Proteins separated by PC1 in the negative direction are downregulated. All highlighted examples are reproducibly quantified across all six samples. STE3 (mating factor a-receptor), MFα (mating factor alpha), RNH70 (exoribonuclease, 5S rRNA maturation), FUI1 (uridine permease), RPL15A (large subunit ribosomal protein), RMT2 (ribosomal protein L12 methylase), and KEX2 (serine protease involved in pro-protein processing) exemplify major factors separating out heat stress from the control due to diminished protein abundance. SBH2 (A, last graph, sh1p-Sss1p-Sbh2p complex component) and SRS2 (B, last graph, DNA helicase) show how PC2 identifies some variability in control and heat stress channels, particularly it seems dependent on variability in control sample 2 and heat stress sample 1 (as is evident in the PCA plot from figure23). Though, as discussed the impact is minimal. Annotations obtain from SGD (http://www.yeastgenome.org/).

Interpretation of Statistically Significant Data

Although the principal component analysis was useful for demonstrating that the majority of variance was due to heat stress and not technical or biological variation (though some existed), the bulk of the relevant data contained within a small space on the PCA loading plot. An alternative means of

plotting the data is to separate out the ratios based on the summed signal to noise of all TMT channels, a proxy for the relative intensity and accuracy of a quantification.

Additionally, statistical measurements can easily be added to these plots, revealing the distribution of significant proteins. Hereafter, "significant" proteins are defined as those proteins whose corrected p-value is less than or equal to 0.01 and whose heat stress to control ratio is greater than or equal to 1.2 (in either up- or downregulated directions). Additionally those protein whose corrected p-value is less than or equal to 0.05 and whose heat stress to control ratio is greater than or equal to 1.5 (in either up- or downregulated directions) were considered significant. This approach was intended to combine the rational use of statistics with a minimum ratio change to ease biological interpretation. Significant proteins are plotted as red triangle and the remainder of the data is plotted as blue circle (Figure 4.25). Highlighted heat stress responsive proteins are listed (top 30 upregulated and top 15 downregulated). The median summed S/N of the significant proteins = 12.2 (log_2 units) vs. 11.7 (log_2 units) for the non-significant proteins, a 1.4 fold greater magnitude. Thus there is no correlation between poor quantification (low S/N) and significantly regulated proteins, as is sometimes observed with mass spectrometry based quantification. An alternative display for the data could be achieved by plotting the –log (p-value) vs. log_2 protein ratios; the displayed method was preferred, however, due to the inclusion of an intensity component. The significantly regulated proteins fall into several classes of biologically relevant processes.

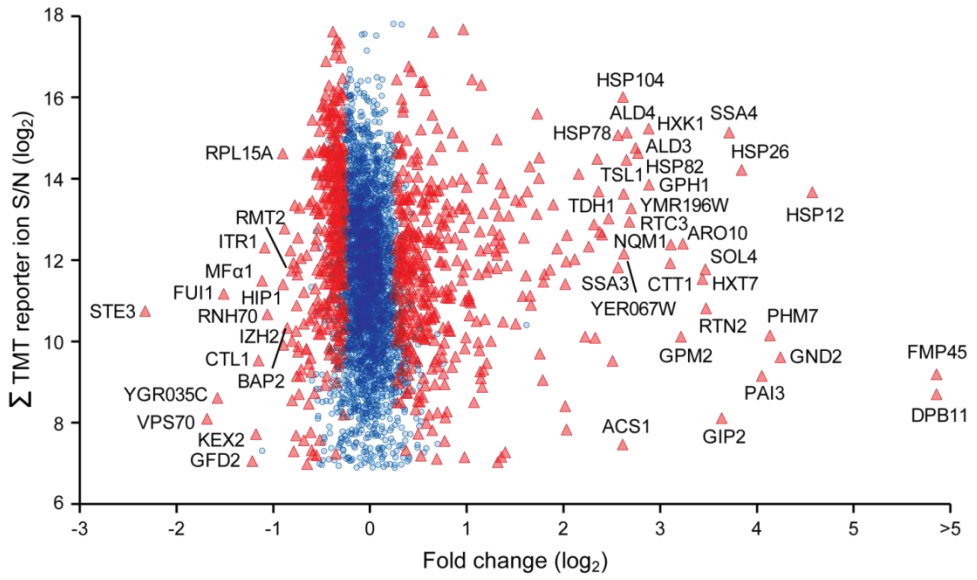


Figure 4.25. Quantification of the yeast heat stress response using statistics. Summed TMT S/N (across all 6 channels) is plotted against heats stress/control ratio sum (129:131)/sum (126:128) for 3, 947 proteins (log_2 ratios). Significant proteins are plotted as red triangles, and the remaining proteins are plotted as blue circles. Those protein changing by 1.2 fold (log +/- ~0.27) or greater, which passed a p-value cutoff of 0.01 (Figure 4.18, B) or by 1.5 fold or greater, which passed a p-value cutoff of 0.05 were considered significant (N = 789). Darkness of color indicates overlapping data points. Summed S/N correlates with data quality, and many of the regulated proteins fall above the median summed S/N (11.7 for all data vs. 12.2, log_2, 40% greater). In contrast to the PCA loading plot, this method of plotting separates out many of the data points in a relevant and presentable manner. The top 30 upregulated and top 15 downregulated proteins are labeled.

To further understand the exact nature of the significant proteins in this data set and their role in the heat stress response, they were separately clustered with the intent of discovering groups of proteins which change by similar magnitudes. Using this sub set of data, 10 obvious clusters were obtained (based on the dendrogram), all of which had unique expression profiles (Figure 4.26, right cluster diagram). These clusters were less obvious when all data was clustered together (Figure 4.26, left cluster diagram). It is possible that the proteins from each cluster may be coregulated by the same cellular machinery or involved in the same pathway; thus, each group of proteins may provide insight into central nodes of protein regulation during the heat stress response (see Figure 4.26 legend for exemplar proteins from each cluster). For example many heat-stress-induced chaperones cluster together at similar magnitudes of expression, as do glycolytic enzymes (Figure 4.26, cluster C). Anabolic enzymes such as those involved in arginine (Figure 4.26, cluster D) and glycogen biosynthesis (Figure

4.26, cluster A) also cluster with one another. Interestingly, components of other pathways have differentially regulated components. In glycogen breakdown, for example, glycogen phosphorylase and the glycogen debranching enzyme fall into different clusters (Figure 4.26, clusters D and B, respectively). Furthermore, members of a single complex were also found to be differentially regulated, such as those in the trehalose synthase complex. The catalytic subunits TPS 1 and 2 were upregulated ~2.5 fold (Figure 4.26, cluster B), whereas the large regulatory component TSL1 was upregulated 6 fold (Figure 4.26, cluster C). This differential expression of catalytic and regulatory components likely has a relevant biological effect. The paradoxical expression of biosynthetic and catabolic pathways seen here is discussed later.

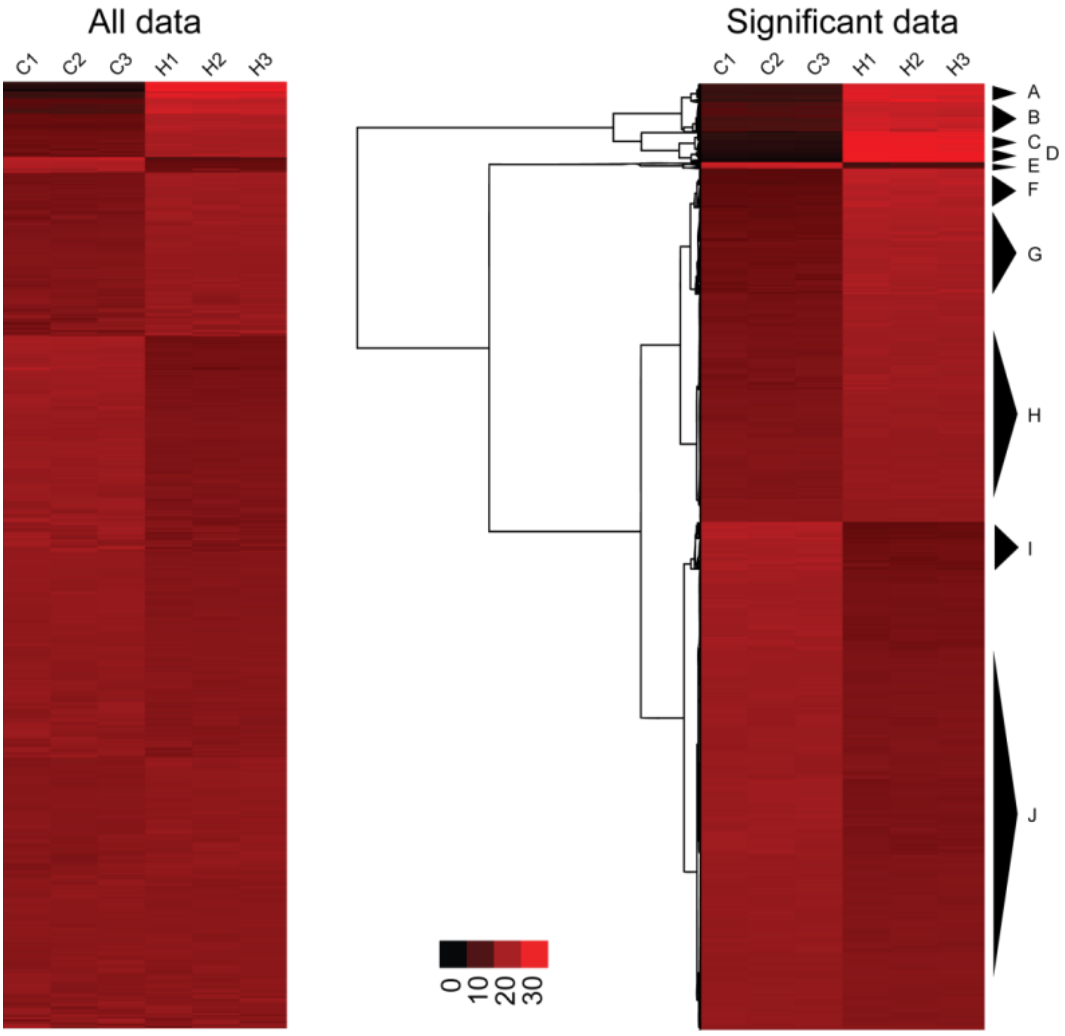


Figure 4.26. Hierarchical clustering of the biological triplicate heat stress. The color bar indicates normalized TMT signal. All proteins (left diagram) and those passing a T-test p-value cutoff of 0.01 (rounded to the nearest hundredth, right diagram) are presented. Proteins and samples were clustered based on Normalized TMT S/N. Using all data combined (left panel) only a small number of distinct clusters were obtained, such as the obvious clustering of upregulated proteins in heat stress at the top of the diagram (protein cluster tree was omitted of simplicity). In contrast, when the data set is limited to those proteins which pass a T-test p-value cutoff of 0.01, many smaller cluster with distinct expression profiles are obtained (right diagram, protein clustering tree shown on the left). Median heat stress/control ratios for each cluster are as follows: (A) 3.37, (B) 2.37, (C) 5.62, (D) 11.09, (E) 0.39, (F) 1.81, (G) 1.49, (H) 1.27, (I) 0.66 and (J) 0.78. Many proteins from each cluster functionally and often evolutionarily related: (A) SSA1, SSE2 (HSP70 chaperones) and GSY2, GLC3 (glycogen synthesis); (B) ARG5, 8 (arginine biosynthesis) and BDH1, 2 (butanediol catabolism, an alternative carbon source); (C) HSP78, 104 (chaperones) and TDH1, PGM2 (glycolysis); (D) SSA4, HSP12 (heat shock proteins) and HXT7, HXK1 (glucose transport and metabolism) (E) STE3, FUI1, RNH70 (no common function); (F) ATG8, PRB1 (vacuolar function/autophagy); (G) IDH1, 2 ACO1, MDH1, LSC2 (TCA cycle); (H) UBP2, 15, UBC7,8 DDI1, FYV10 (ubiquitin proteasome system); (I) AGP1, SAM3, HIP1(amino acid permeases), RPL14, 15, 16 (large ribosome subunit), and REX3,4, NSA2 (rRNA processing); (J) RPL 25, 28, 35 (additional large ribosome subunit), RPS0A, 11, 12, 15 (small ribosome subunit), CDC33, TIF4631 (translation initiation factors) and YEF3, EFT2 (translation elongation factors). With the use of statistical significance, more subtle protein clusters become relevant, and the contributions of both obvious and no-obvious proteins to heat stress can be analyzed.

Extracted Biology and its Role in the Heat Stress Response

To help narrow down the relevant biology involved in the heat stress response, significant proteins were analyzed for enriched gene ontology categories for cellular compartment and biological process. Gene ontology revealed a broad trend in the significantly upregulated and significantly downregulated data, consistent with those found through PCA. In general upregulated proteins are involved in metabolism and aspects of protein maintenance (folding and degradation). Downregulated proteins comprise primarily those proteins involved in ribosomal processes. Downregulated proteins were enriched with both nuclear (nuclear lumen, more specifically nucleolus, 4 fold enrichment) and cytosolic cellular locations (2 fold enrichment). Generally the nuclear proteins were part of rRNA processing complexes (e.g. small subunit processome, 5 fold enrichment), whereas cytosolic components were ribosome complexes themselves (e.g. large subunit fold ~3 fold enrichment). There was a slight, albeit just below statistically significant, enrichment of cytosolic protein in the upregulated data set (~1.6 fold enrichment). There was, however, statistically relevant enrichments for the mitochondrial matrix and mitochondrial membrane (~2 fold enrichment), the plasma membrane (~4 fold enrichment), the trehalose synthase complex (~10 fold enrichment), and peroxisomes (~3 fold enrichment). Enrichment of these cellular compartments is consistent with their function in metabolism and protein maintenance. GO cellular compartment data is summarized in Figure 4.27.

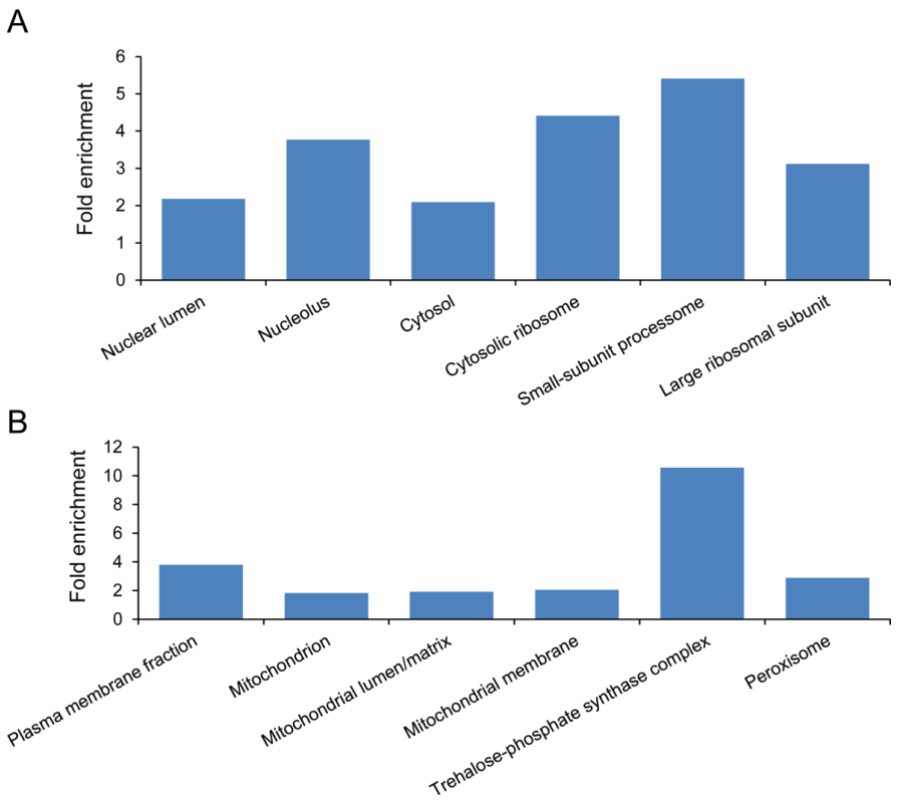


Figure 4.27. Extracted gene ontology terms for cellular compartment, from the group of significant proteins. A small subset of related GO terms are presented and plotted against their fold enrichment. (A) Downregulated terms were enriched with both cytosolic and nuclear proteins (2 fold each), with a larger (4 fold) enrichment for the nucleolus. Consistent with this observation, complexes of ribosomal proteins and rRNA/ribosomal processing machinery were enriched. (B) Upregulated proteins were enriched with membrane associated, mitochondrial (membrane and co-occurring term lumen/matrix) and peroxisome proteins. Additionally the trehalose-phosphate synthase complex was greatly enriched (~10 fold), which is consistent with the observation that trehalose is an important reserve carbon source and structural component during environmental stress response. In general highly similar categories were omitted (high annotation overlap and similar enrichment fold change).

Although the cellular distributions proved to be interesting, of greater importance may be the biological functions of the proteins regulated by heat stress. Protein flux and proper maintenance of proteins were significant; Protein translation (~ 2 fold enrichment) and degradation (including proteasomal degradation, ~2 fold enrichment, and autophagy, ~6 fold enrichment), and protein folding (~8 fold enrichment) were overrepresented. Translation aspects were extracted from downregulated proteins, whereas the other aspects were extracted from upregulated proteins. These aspects of translation (generally enriched 3 fold or greater) include ribosome biogenesis (rRNA processing and

ribosome assembly) non-coding RNA metabolism (rRNA and tRNA), and the nuclear export into and localization of ribosomes within the cytoplasm. Other significantly enriched biological process categories, involved in various aspects of metabolism, were extracted from the upregulated data set.

These processes (generally catabolic in nature) are involved in ATP production (aerobic respiration, ~3 fold enrichment, and TCA cycle, ~4 fold enrichment). Various categories of carbohydrate catabolism (~3 fold enrichment), including traditional pathways such as glycolysis (hexose catabolism, 3 ~fold enrichment) were overrepresented. The utilization of alternative carbon sources was also present, including pentose catabolism (~7 fold enrichment) and alcohol catabolism (~3 fold enrichment). Interestingly the biosynthesis of trehalose and glycogen are enriched in the upregulated protein data set (~8 and ~7 fold enticement, respectively), counterintuitive to the narrative of carbohydrate utilization. It has been well documented, however, that these sugars are important for various aspects of the stress response[27, 28]. Lipid catabolism (~4 fold enrichment) and amino acid catabolism (~4 fold enrichment) were present. Finally, metabolically related categories of NAD and NADP metabolism (~3 fold enrichment) and cellular redox homeostasis (~4 fold enrichment) were significantly extracted from the upregulated data. GO biological process data is summarized in Figure 4.28.

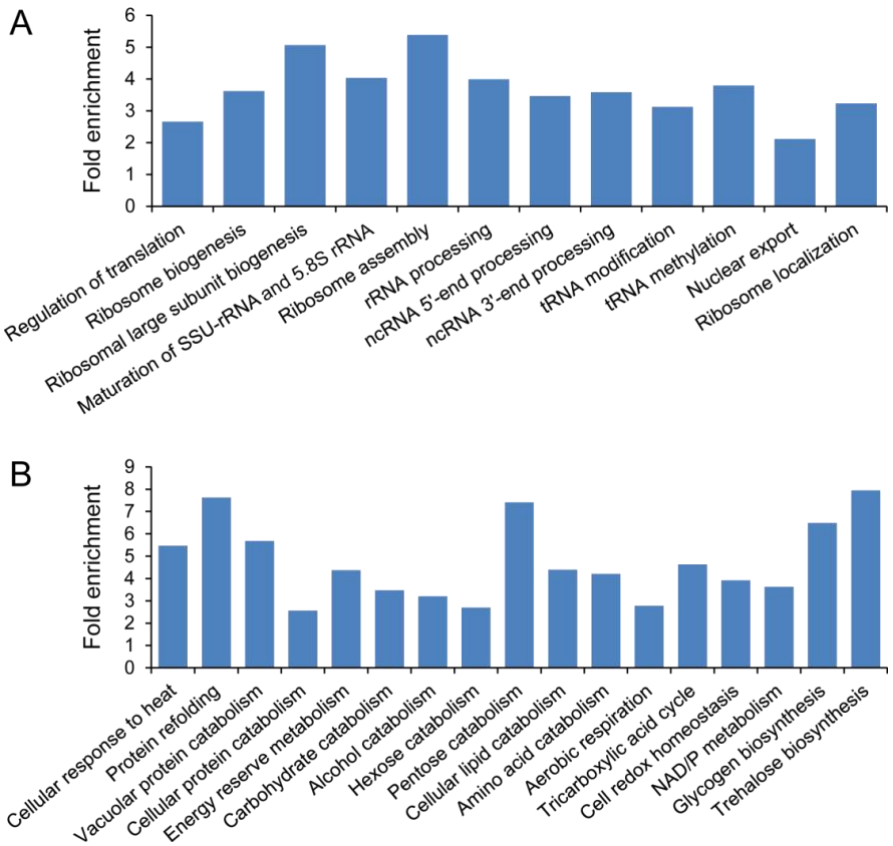


Figure 4.28. Extracted gene ontology terms for biological process, from the group of significant proteins. (A) Downregulated protein GO terms are highly enriched with various facets of ribosome production and protein translation, from the ribosomal proteins themselves, to rRNA/ncRNA processing, tRNA modification (methylation) and nuclear export and cytosolic localization of ribosomal components. (B) Upregulated proteins are enriched with a wealth of GO categories, many of which are associated with metabolism, particularly catabolic processes. As expected cellular response to heat and protein folding are enriched (~5.5 and ~8 fold, respectively). Protein catabolism and vacuolar process are enriched (~5 and ~2 fold, respectively), indicating an increase in proteasomal and autophagosomal degradation. Carbohydrate, lipid and energy reserve metabolism categories are highly enriched, including those involved in alcohol, pentose and amino acid catabolism (~3-8 fold). Aerobic respiration, including TCA cycle proteins and those involved in NAD/NADP metabolism are enriched in the upregulated data set (~5 fold). Interestingly many redox-related proteins are also enriched by a similar magnitude. Finally, the reserve sugars of glycogen and trehalose are highly enriched. In general highly similar categories were omitted (high annotation overlap and similar enrichment fold change).

It is clear from the GO categories that a diverse array of processes occur upon heat stress. We can break those mentioned into two main categories; Protein homeostasis (translation, maintenance and degradation) and nutrient metabolism (catabolic and anabolic processes, including secondary metabolic functions of NAD and NAPH metabolism and redox homeostasis). Each category and its role in the heat stress response are discussed.

Probably the most recognized response to heat stress is the upregulation of molecular chaperones, as a result of protein unfolding or misfolding. Indeed many of the most highly upregulated proteins are HSP chaperones (Figure 4.26 clusters C and D, e.g. HSP26, HSP42, HSP82, HSP70 family, HSP104). Many of these proteins are ATP dependent chaperones, and thus the accumulation of a large number of improperly folded proteins would require a large amount of energy to alleviate. This energy requirement has been suggested to be at least partially responsible for the increased metabolic activity observed in heat stress (discussed below[23]). A protein interaction network of the HSP70 family of chaperones (constructed using Genemania) reveals a highly interconnected network with apparent coregulation (Figure 4.29). Upregulated proteins in this network include additional HSPs and other chaperones, including HSP42, SIS1, SGT2, APJ1 and others. Interestingly the nucleotide exchange factor for SSA1 (and ATP dependent HSP70 family member), FES1[29], was coregulated with the upregulation of SSA1 in this network. SSB1 and SSB2, HSP70 family members which are known to be downregulated upon heat stress[30], are indeed found downregulated in this network. Interestingly, these proteins have been proposed to be associated with active ribosomes[31] and nascent polypeptide chains. Other ribosome-associated proteins are found in this network as well (e.g. TMA46). As they function with active ribosomes, it is possible that proteins such as SSB1 and SSB2 are co-regulated (downregulated) with ribosomal proteins (discussed below).

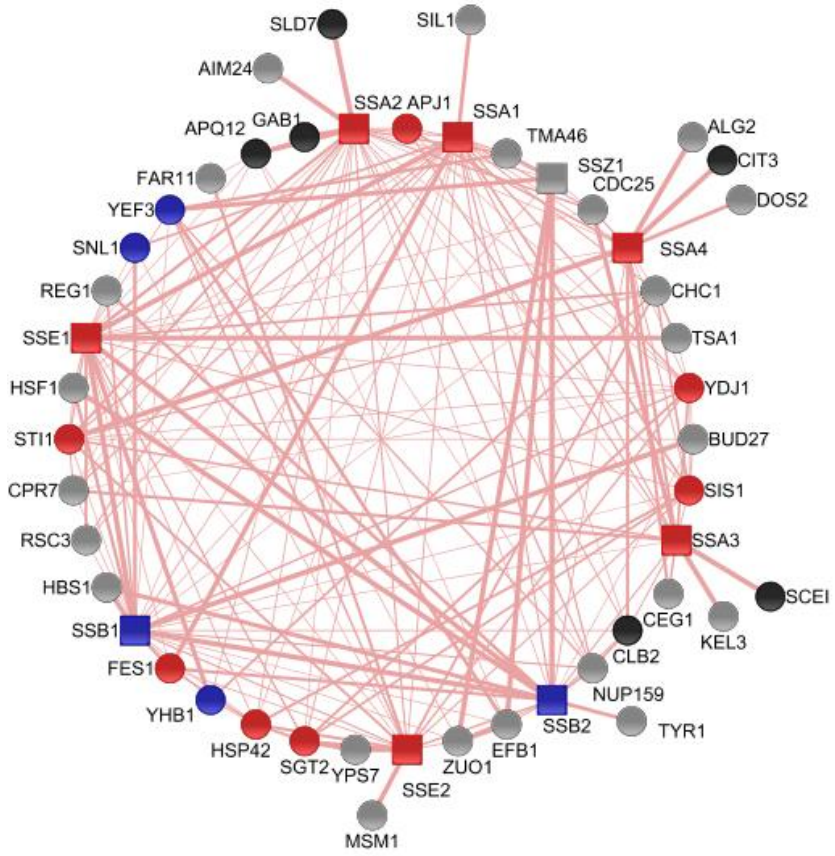


Figure 4.29. Physical interaction network of the HSP70 family of chaperones. Genemania was queried for the top 40 physical interactions (circles) of the HSP70 family of conserved chaperone proteins (SSA 1-4, SSB 1-2, SSE 1-2, and SSZ1, squares). Edges represent interactions and their weight was assigned automatically by Genemania based on the literature. Nodes were colored according to their expression: black, not quantified; gray, no significant change upon heat stress; red, significantly upregulated upon heat stress; blue, significantly downregulated upon heat stress. Very strong (based on literature evidence) connections were observed in this network, as well as a high level of interconnectivity. Besides the queried genes SSB1 and SSB2, other network components are downregulated: YHB1, nitric oxide reductases, consumes NO. As discussed below, NO production (from arginine) may be an important stress adaptation, thus reducing the consumption rate of NO may be beneficial in the heat stress response; YEF3 (translation elongation) and SNL1 (ribosome associated, various functions) are in involved in the positive regulation of translation, and their downregulation is consistent with other observed decreases in ribosome machinery. Many upregulated proteins from the network are notable: Additional chaperones are present, including HSP42 and ADJ1; STI1, SGT2, SIS1, and YDJ1 are co-chaperones; FES1 is the SSA1 nucleotide exchange factor, and positively contributed to its activity. Interestingly, HSF1 itself is unchanged on the protein level, though it is known its mRNA levels are affect by heat stress, suggesting some level of posttranscriptional control (it is known to be hyperphosphorylated in many stressed states) or transient expression.

Ribosomal proteins and other machinery involved in protein translation represented the main group of downregulated proteins in this analysis. Both large (e.g. RPL12B and RPL15A) and small (e.g. RPS4B and RPS11B) subunit protein components were found to be downregulated. Rap1 protein expression, a transcription factor known to regulate ribosomal protein gene expression[32], was

unchanged, however. Interestingly RAP1 is also implicated in glycolytic enzyme expression[33].This result suggests that either a RAP1-independent mechanisms exists, or an alteration in RAP1 DNA binding or protein interactions, perhaps through posttranslational modifications, occurs in response to heat stress. Other non-structural proteins involved in the regulation of ribosome assembly, tRNA modification and rRNA processing were also significantly downregulated in this analysis. Intriguingly, RMT2, an arginine methylase, for which RPL12 (highlighted above) is a substrate[34], is among this group of proteins; this methylation event (likely diminished in heat stress) may be a positive regulator of ribosome assembly, based upon similar modifications of other ribosomal proteins[34].The downregulation of NSA2, a protein involved in rRNA processing, and TRM2, a tRNA methylase, demonstrate that non-coding RNA biology is also important in the heat stress response. In fact, tRNA methylation has been shown to positively regulate the rate of translation[35].

The end result of these regulatory events is likely a large reduction in ATP consumption from reduced protein synthesis, which may be crucial for handling the increased energy load brought on by high chaperone activity. It has also been suggested, however, that the diminished expression of protein synthesis machinery may be a response to aggregation of ribosome assembly intermediates during stress[36]. These aggregates compete for chaperones[36] inhibiting the proper folding of other required proteins. In most cases, the components discussed were regulated at magnitudes <2 fold (though significant), much less than many upregulated proteins, which may suggest there is an intricate balance between the necessity for stress related translation and the deleterious effects translation in general (or aggregation of involved proteins) may have during the stress response. Additionally it supports the notion that subtle changes are biologically relevant, and the use of statistics-based proteomic methods was required to understand their expression. Translation is one factor in protein flux, the other factor is degradation.

Both proteins from the ubiquitin proteasome system and the autophagy pathways (as well as associate vacuolar processes) were significantly upregulated in heat stress, including for example UBC8, FYV10, and ATG8. UBC8 and FVY10 play a role in the ubiquitin mediated degradation of fructose-1, 6-bisphosphatase, thereby inhibiting gluconeogenesis[37, 38]. ATG8 plays a role in phagopore expansion during autophagosome formation, and its relative expression has been shown to be a determinant of autophagosomal size[39]. Protein degradation through the ubiquitin system may be primarily useful for clearing misfolded proteins that results from the sudden temperature shift as well as the degradation of regulatory proteins (as highlighted above). It is likely, however, that vacuolar protein degradation through autophagy is a general response to the nutrient depleted state[40] which heat stress mimics (similar to stationary phase[23]). It is known, for example, that the TOR pathway inhibits (both directly and through SCH9) autophagy[41], and that TOR signaling itself is inhibited during the heat stress response[42].

The majority of the non-chaperone upregulated proteins consist of a variety of metabolic proteins, generally those involved in catabolic processes. A large number of the glycolysis pathway enzymes are upregulated (HXK1, GLK1, and PGM2). In coordination with glycolysis, proteins involved in aerobic metabolism, such as those in the TCA cycle are upregulated. Interestingly, many proteins upregulated by glucose limitation are also upregulated during heat stress, despite the presence of high glucose in the growth medium and the upregulation of glucose utilization pathways. For example, the hexose transporter HXT7 is highly upregulated. Additionally, proteins thought to be expressed during stationary phase (e.g. TDH1), some for the utilization of alternative carbon sources (e.g. ethanol catabolism, ACS1), are regulated in heat stress. Consistent with the idea that stress induces a starvation response, many of the regulated proteins in heat stress were also found to be regulated in yeast upon rapamycin treatment[43]. Alternatively, aspects of the starvation response may itself be part of the stress response.

Peroxisomal proteins involved in NAD metabolism (PNC1, NAD salvage) and those in the pentose phosphate pathway (SOL4, GND2, NADP metabolism) are upregulated. Non-peroxisomal proteins involved in *de novo* NAD production (BNA 1, 5 and 6) were also upregulated. These results are not surprising in light of the importance of NAD and NADP play in metabolism and redox reactions. The upregulated proteins TSA2, GRX1, and GRX2, are involved in redox homeostasis more specifically and may be useful at managing the increase in reactive oxygen species (ROS) which results from increased aerobic respiration. Indeed many redox proteins regulated by heat stress are also involved in DNA replication stress[44], a potential outcome of increased ROS generation. These proteins may also be important for maintaining a proper redox balance for oxidative phosphorylation.

In many cases the exemplified metabolism-related proteins show a stress specific pattern of isoform expression (e.g. TDH1 but not 2 and 3 are upregulated, Figure 4.30), which may be an indication that they have unique properties compared to other isoforms. Further research assessing the point of isoform regulation, whether due to differential transcription, translation, or protein stability, would be extremely useful for explaining these stress specific observations.

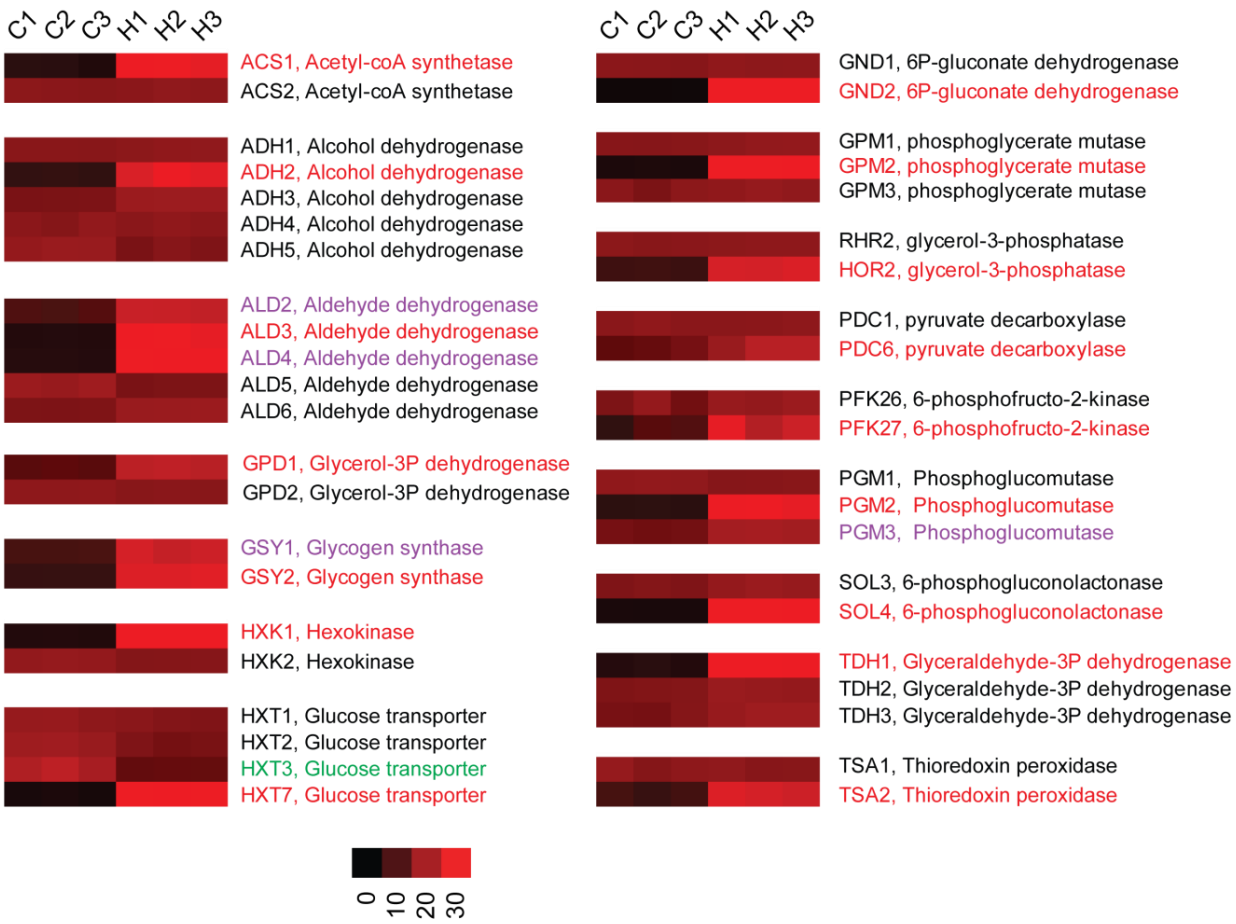


Figure 4.30. Differential regulation of protein isoforms upon heat stress. A large number of proteins with functional isoforms were identified and quantified in this analysis. Heat maps of the relative TMT S/N across all control and heat stress samples are presented. The isoform displaying the greatest response to heat stress is highlighted in red, while other isoforms that are regulated are highlighted in purple. In the case of the glucose transporters, HXT3 was significantly downregulated, and labeled in green. Interestingly, even when many isoforms are present, often only one of the isoforms is upregulated upon stress. As many metabolic proteins are regulated by heat stress, it is likely that the various isoforms listed here which are upregulated upon heat stress participate in those processes, such as alternative carbon source utilization, protein catabolism and glycogen/trehalose accumulation, and redox homeostasis. All quantifications were double checked using only unique peptides (no database redundancy), to ensure sequence redundancy was not affecting the quantification. No differences were found when using only unique peptides.

Paradoxical Protein Regulation in the Heat Stress Response

Interestingly, as was reported using a genomics approach[23], many biosynthetic and catabolic process of the same biomolecules are simultaneously and paradoxically upregulated during heat stress. For examples both trehalose and glycogen biosynthetic (e.g. TSL and GSY2) and utilization pathways (e.g. ATH1 and GPH1) are upregulated. Protein interaction networks for these pathways are displayed in

figures 4.31 and 4.32. These sugars are both important in reserve energy metabolism, and also may play a protectant role in membrane biology[27, 28].

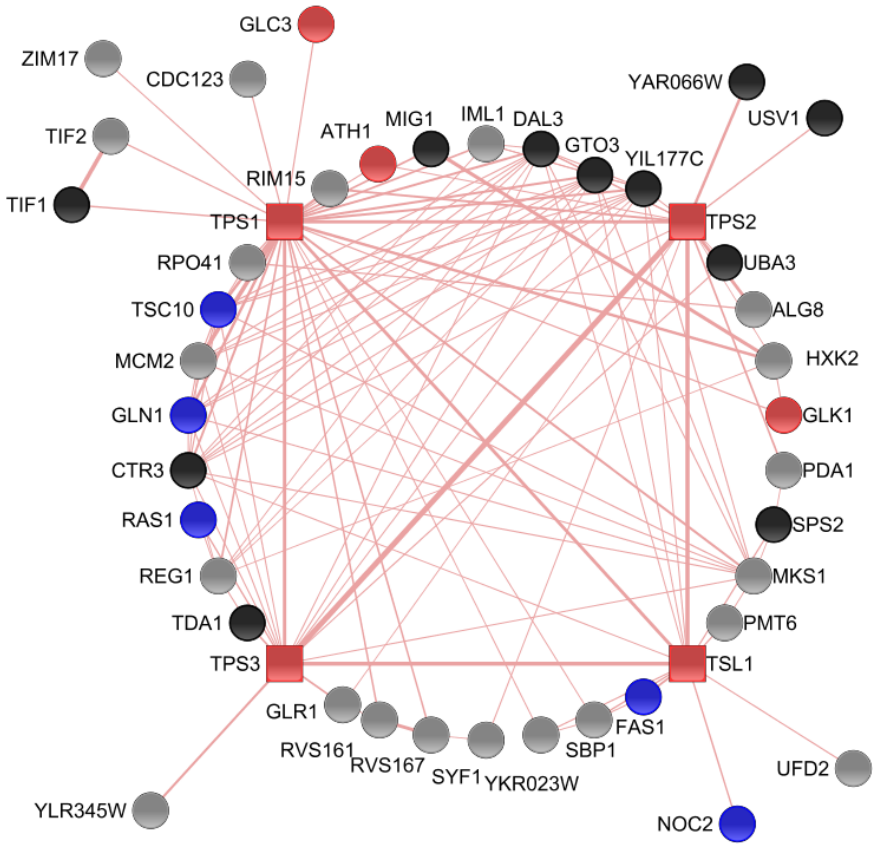


Figure 4.31. Physical interaction network of the trehalose synthase complex. Genemania was queried for the top 40 physical interactions (circles) of the trehalose synthase complex (TSL1, TPS1-3, squares). Edges represent interactions and their weight was automatically assigned by Genemania based on the literature. Nodes were colored according to their expression: black, not quantified; gray, no significant change upon heat stress; red, significantly upregulated upon heat stress; blue, significantly downregulated upon heat stress. Many of the upregulated components are relevant to heat stress: All components of the complex were upregulated, though at different magnitudes as discussed above; ATH1 (acid trehalase), as well as NTH1 (neutral trehalase, not present in network) were upregulated upon heat stress and are required for the degradation and utilization of trehalose (paradoxically regulated with the trehalose synthase complex). The interactions of TPS1 (regulatory subunit) and GLK1 (glucokinase) may be important for regulating trehalose synthase, as glucose-6-phsophate is the substrate of the trehalose synthase complex. GLC3, glycogen branching enzyme, may provide a connection between stress induced glycogen and trehalose synthesis processes. Downregulated components are noteworthy as well: RAS1, involved in cAMP/PKA signaling; FAS1, fatty acid synthesis; NOC2 is involved in intranuclear ribosome transport, and may help connect metabolic status to the ribosome machinery.

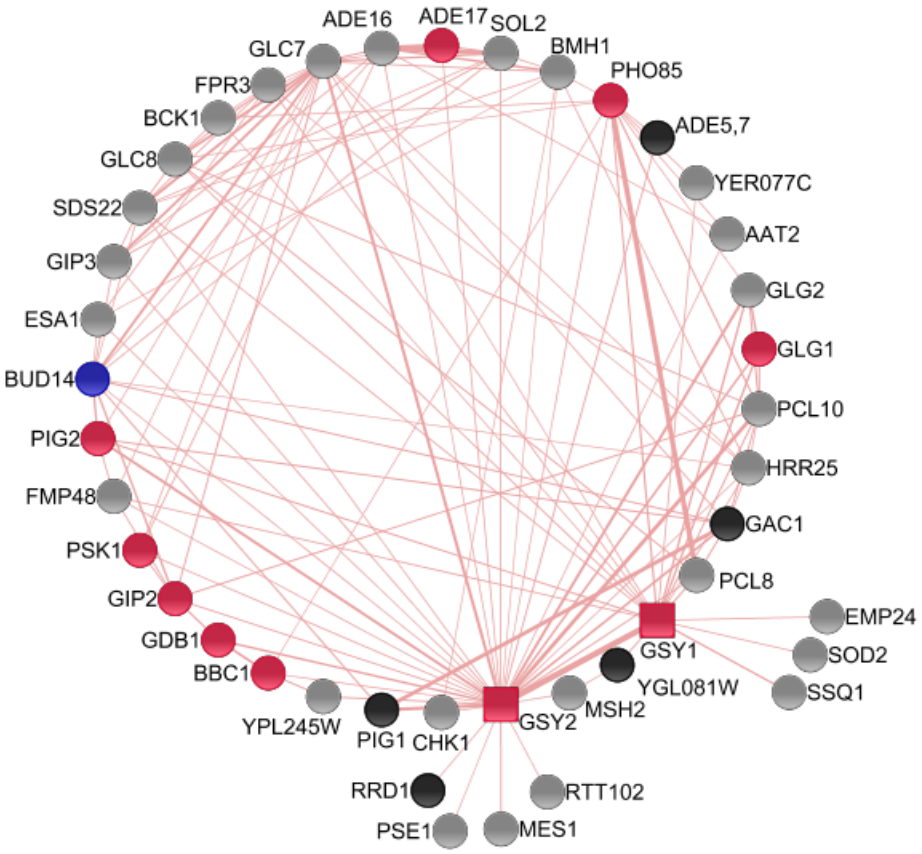


Figure 4.32. Physical interaction network of the two glycogen synthases. Genemania was queried for the top 40 physical interactions (circles) of the two glycogen synthase isoforms quantified in this experiment (GSY1 and GSY2, squares). Edges represent interactions and their weight was assigned automatically by Genemania based on the literature. Nodes were colored according to their expression: black, not quantified; gray, no significant change upon heat stress; red, significantly upregulated upon heat stress; blue, significantly downregulated upon heat stress. Both isoforms are regulated by heat stress, though GSY2 is more highly induced, also contains more interacting partners (in the context of these results), and a stronger association amongst some of the common ones, which may be relevant in its function or regulation during heat stress. BUD14, the only downregulated component, is a regulatory component of GLC7 phosphatase activity (a network component, not regulated). GLC7 regulates a diverse number of processes including glycogen metabolism, glucose repression, and cell wall organization, all of which are important in the stress response. Many upregulated network components are noteworthy: GIP2 and PIG2, additional regulatory subunits of the GLC7 phosphatase, similar to GAC1 (network component, not quantified). GAC1 causes GSY2 and GLC7 interaction leading to glycogen accumulation, and thus GIP2 and PIG2 may have similar functions; GLG1, glycogenin glucosyltransferase, an initiator of glycogen synthesis; PSK1, poorly characterized nutrient sensing (by unknown mechanism) kinase; GDB1, glycogen debranching enzyme; PHO85, a cyclin dependent kinase involved in nutrient sensing, inhibits GSY2 (when bound to PCL 8/10, also network components, unchanged. Both positive and negative signals of glycogen accumulation are contained within this network, further complicating the perplexing nature of glycogen metabolism during the heat stress response.

It was previously observed that both biosynthetic and catabolic pathways of these sugars were regulated at the transcript level as well[23]. The author suggested that posttranslational control of one

branch over the other allowed yeast to contribute to or draw from the reserve pool to manage the protectant and reserve metabolite nature of each branch. These data show that both branches are simultaneously upregulated at the protein level, consistent with the suggestion that posttranslational regulation is responsible for controlling which pathway is active. High expression levels of each branch would allow for the rapid flux of substrates through whichever pathway is active, as increasing an enzyme's concentrations raises the Vmax of a reaction. The glycogen branching and synthase enzymes were upregulated ~3.5 fold, whereas the glycogen phosphorylase and debranching enzymes were upregulated ~7 and ~2 fold respectively. Although the phosphorylase enzyme is the most highly upregulated, the average expression between the biosynthetic and utilization pathways are similar. Each branch may be in constant flux depending on the cell's need for energy vs. structural protection, and this type of regulatory scheme would permit a rapid response to either need. Additional levels of regulation may also be involved, such as the cellular locations of these enzymes.

Other seemingly paradoxical pathways were also found to be coregulated in heat stress. Proteins involved in amino acid catabolism and amino acid synthesis were found to be upregulated simultaneously. A closer look at the pathways involved, however, begins to reveal a potentially interesting biological explanation for such regulation. The catabolic process mentioned is that of the Ehrlich pathway, whereas the anabolic process is the biosynthesis of arginine. Upstream components of the Ehrlich pathway such as the aminotransferase ARO9 and decarboxylase ARO10, as well as downstream components such aldehyde and alcohol dehydrogenases (e.g. ALD3, and ADH2) were found to be upregulated upon heat stress. The Ehrlich pathway has traditionally been of interest in the brewing industry, due to the production of fusel acids/alcohols from amino acids, which affect the product quality[45]**.** It has been demonstrated that improper temperature control during the brewing process increases the production of fusel alcohols[46], supporting its role in heat stress resistance. This pathway is used by yeast for the assimilation of amino acid nitrogen sources, via transamination of α-

keto glutarate to glutamate from catabolized amino acids. Additionally the Ehrlich pathway may be important for maintaining the NAD+/NADH balance (dependent upon aldehyde vs. alcohol dehydrogenase branches of the pathway[45]), which may be important in heat stress resistance (as exemplified by the enrichment of NAD and NADP go terms, Figure 4.28). Arginine, however, is not a substrate for this pathway; conversely, the pathway product glutamate is a substrate for the production of arginine, thus offering a biologically sound solution to the apparent metabolic paradox.

Many proteins involved in the biosynthesis of arginine were found to be upregulated. These include ARG3, ARG5, ARG8 and CPA1, proteins involved in the production of citrulline, an arginine precursor. Traditionally, these proteins are thought to be repressed in the presence of arginine, both transcriptionally and transitionally[47, 48]. Due to the fact that arginine is indeed present in the growth medium (synthetic complete media), evidently an additional level of regulation is occurring upon heat stress. Proteins involved in arginine catabolism and arginine to proline conversion, however, were not regulated in heat stress (e.g. CAR1, 2 and PRO3). This result suggests that it is arginine itself and not its function as an intermediate for proline synthesis, which may be relevant. The intracellular concentration of arginine may play a key role in heat stress resistance, either directly as compound (perhaps due to its charged nature), its role in cellular stress pathways, or through its incorporation into arginine rich proteins.

Arginine was shown to directly reduce protein aggregation in a concentration dependent manner (within the physiological range of concentrations), through its suppression of intermolecular interactions among aggregation-prone molecules[49]. Additionally arginine has been found to be important for resistance to oxidative stress[50]. The authors suggest that that nitric oxide (NO), as previously observed in other stress conditions[51], is important for oxidative stress resistance. Furthermore they suggest that arginine's conversion to NO via a yet to be identified nitric oxide synthase (NOS) is responsible for the stress protective effects of arginine, linking arginine synthesis and

stress protection. It has also been suggested that the generation of reactive oxygen species during the heat stress response (perhaps due to the aforementioned increase in catabolic processes) may be partially responsible for the deleterious effects of heat stress[52, 53]. Thus in a similar way, arginine may be protective during heat stress as with oxidative stress. Indeed recently, it was demonstrated in heat stress directly that NO production from arginine conferred a stress protectant effect in yeast[54]. This analysis has demonstrated that proteins which are seemingly contradictorily regulated in large scale data sets begin to make sense when the specific pathways involved are dissected, and the biology is interpreted.

A Proteomic Time Course Analysis Reveals Dynamically Regulated Heat Stress Responses in Yeast

Though the statistical analysis of a single time point in the yeast heat stress response unraveled a large number of regulated protein responses, it is likely that different proteins are responsible for the adaptive mechanisms of the heat stress response at different times during duration of the stress; others may be required throughout the entire stress event. Yeast display no growth delay during heat stress, which is in contrast to other environmental stresses such hyperosmotic stress[25] and oxidative stress (data not shown). Hence, there is no obvious period of adaptation from which to guide a temporal analysis. Therefore, a number of time points are required to accurately capture the stress response, a task particularly suited for TMT. This data may also indicate that yeast are particularly well adapted to environmental fluctuations in temperature. Indeed with the exception of nutrient availability, heat stress may be the most frequent stress experienced by yeast in the natural environment. To further understand the nature of the heat stress response in yeast, a time course of 0, 30, 60, 90, 120, and 240 minutes after heat stress (37 °C) was obtained, time points which encompass the majority of logarithmic growth at 37 °C (Figure 4.33). Here I demonstrate the use of TMT in a time course analysis, highlight

relevant methods for analyzing the data, and remark on the different programs of regulation during heat stress.

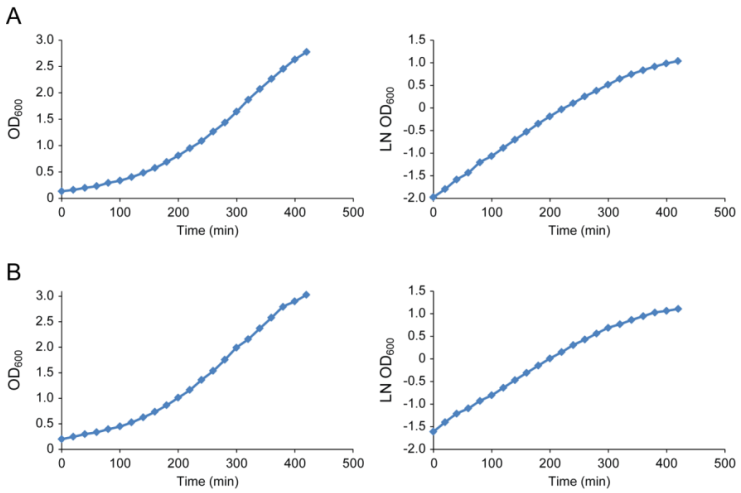


Figure 4.33. Growth curve of unstressed (A) and heat stressed yeast (B). Both unstressed and stressed yeast grow at a normal rate, with doubling times of approximately 90 minutes. Logarithmic growth occurred from an OD_{600} of approximately 0.2 to 2.0. Stress experiments were always performed in mid log phase. Interestingly heat stressed yeast show no growth delays as compared to other stresses, such as salt stress and oxidative stress.

Quantification Statistics of the Heat Stress Time Points

By 30 minutes of heat stress, already many proteins were found to be regulated (>2 standard deviations, Figure 4.34, A). The number of proteins regulated at a given time point remained fairly constant throughout the time course, though the magnitudes of regulation (fold change) tended to increase over time; this increase in the fold change was accompanied by an increase in the data set variance (Figure 4.34, B). As a result a plus or minus two standard deviation cutoff for relevance became more stringent over time, explaining the similar numbers of regulated proteins between time points, and highlights the need for additional replicate analyses. A drop in the magnitude of protein change between the 90 and 120 minute time points, which is regained at the 240 time point, may be indicative of transiently regulated and late acting proteins or experimental noise; the 90-120-240 minute transition may reveal proteins which begin to fall below relevance, while others begin to gain relevance. Further supporting this notion, many proteins were found to be regulated at only one or two time points (Figure

4.34, C). A number of proteins, though, were found to be regulated throughout all five stress time points, perhaps suggesting their constant expression is required for higher temperature adaptation. Though each individual time point may reveal an interesting result, the trend of protein expression over all time points is likely more relevant to this analysis.

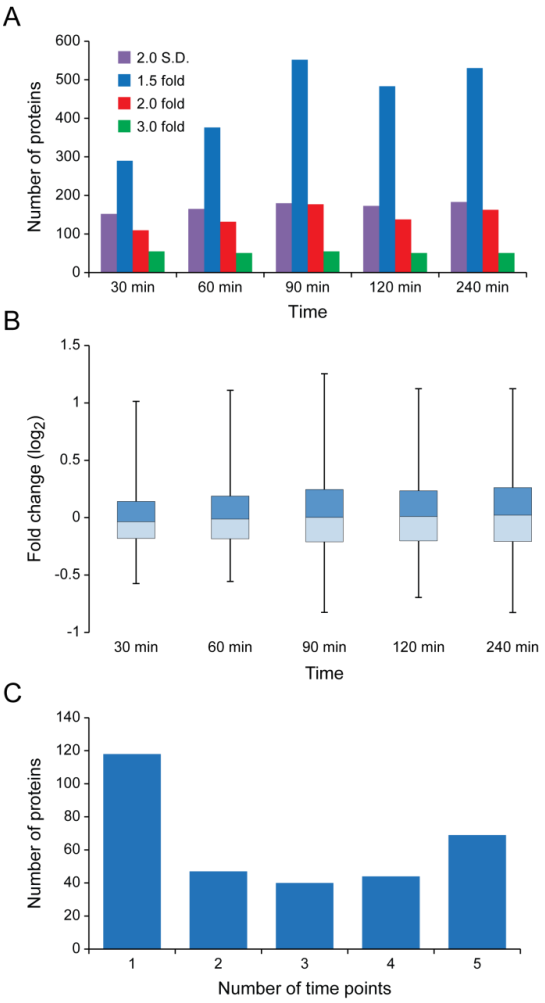


Figure 4.34. Heat stress time course data set statistics. (A) Histograms of regulated proteins and protein expression magnitudes for each time points. Hundreds or proteins were regulated at each time point during the heat stress response, defined by a fold change of 2 or more standard deviations. Generally the magnitude of protein changes, defined by the number of proteins demonstrating the given fold change, increased as time progressed. It is interesting, however, that protein expression patterns demonstrate a transient drop in magnitude at the 120 minute time point. This behavior may be due to the convergence of temporally regulated proteins; moreover, some regulated proteins may be returning to steady state levels between 90 and 120 minutes, while others begin or continue progressing away from steady state levels at the 240 minute time point. Replicate analyses would confirm this hypothesis. Of importance, many proteins are already regulated at the 30 minute point, and are thus likely needed for the early stress response. (B) Heat stress/control ratios show similar distributions between time points, standard box plots with min/max values capped at the 2nd and 98th percentiles for the sake of presentation. (C) Histograms of regulated proteins vs. number of time points. Many proteins are regulated a just a few time points, suggesting they may be transiently regulated. Others, which are regulated in four or five of the time points, may be of constant need during heat stress.

Dimensionality Reduction of the Heat Stress Time Course Data Reveals Groups of Temporally Regulated Proteins

Often in biology, many simultaneous processes function to cooperatively regulate a particular pathway or stimuli response, rendering a thorough definition of the system challenging. Particularly with large-scale quantitative approaches such as genomics and TMT based MS-multiplexing, many variables

exist within a data set, creating a need for dimensionality reduction of these variables. Once the system is divided into a more manageable number of parts, interesting patterns in the data often emerge. To permit such dimensionality reduction, two comparable methods were applied to this data set: principal component analysis (PCA), and non-negative matrix factorization (NMF). As discussed, both methods break the protein expression data into a variety of matrices which allow for the explanation of data set variance. NMF contains additional benefits which may be suited to quantitative proteomics, such as the ready interpretation of the extracted components. The use of NMF in the analysis of this data set is the first application of NMF in proteomics. We generally expected up to four possible groups to exist within the data: those containing proteins which were up or downregulated throughout the time course, and those containing transiently up and downregulated proteins. Generally this assumption held true.

Prior to executing PCA and NMF, the data was hierarchically clustered to observe the relationship between each time point. This method was an important first glimpse at the success of the experiment. The time points clustered in primarily in the expected pattern (Figure 4.35, A): The stressed time points clustered away from the unstressed control; the closest stressed time points clustered together (e.g. 90 and 120 min). Interestingly the 240 minute time point was related to both the unstressed control and the other group of stressed time points, again supporting the notion that transiently regulated proteins exist in the data set. With this knowledge PCA was first performed to assess how many relevant components existed in the data, and to what degree each component explained the data set variance (Figure 4.35, B).

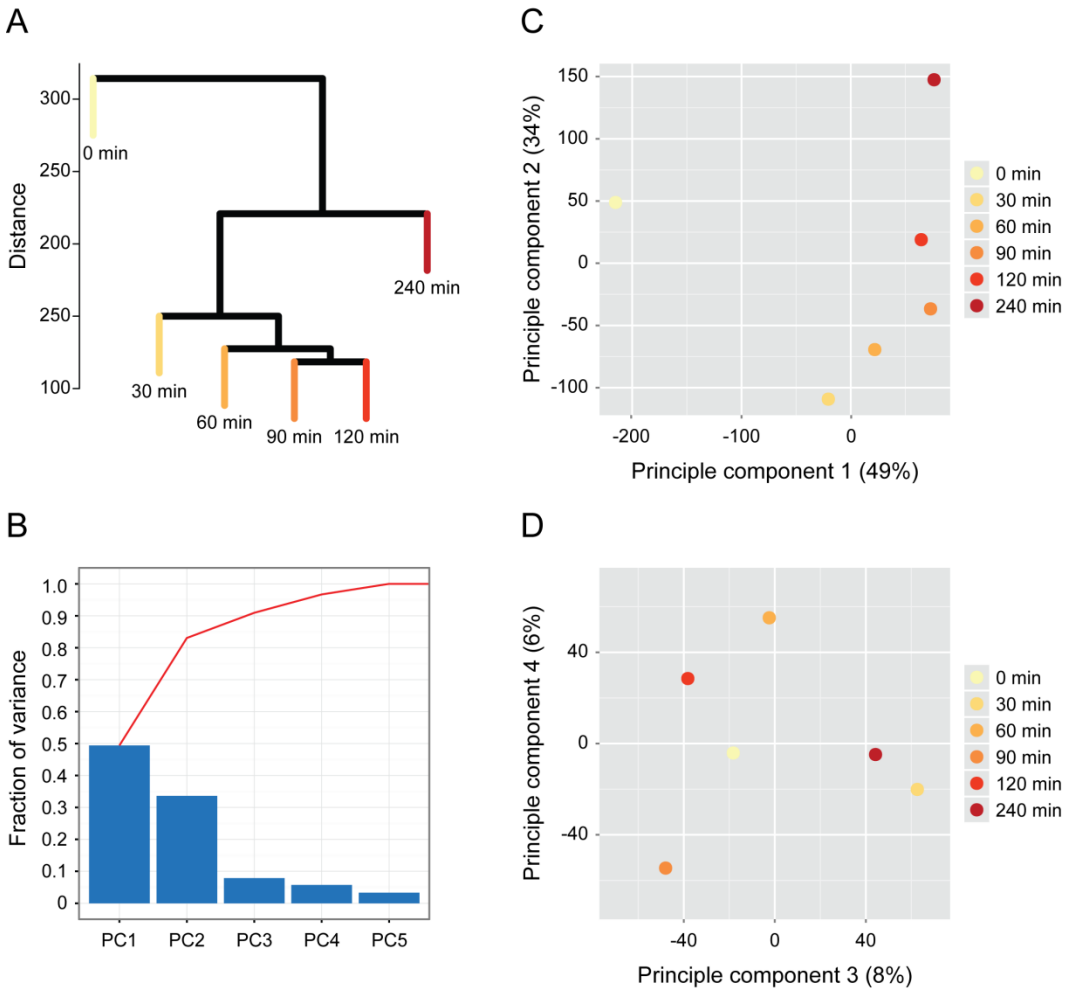


Figure 4.35. Principal component analysis of time course data reveals two primary components. (A) Hierarchical data clustering revealed the expected relationships amongst data points. Each array was composed of the relative TMT S/N across all proteins (so that each protein array sums to 100% across the 6 samples), and linkage was carried out using the Euclidian distance similarity metric and centroid method (though other methods were extremely comparable). Interestingly the 240 time point clustered intermediately between the control and the groups of other stressed time points. (B) Two main components were observed after PCA, which explained nearly 90% of the data set variance. (C) The distribution of time points within each of the first two components reveals a biological explanation of the components: the first component separates the stressed time points from the non-stressed control, likely representing upregulated (along the positive component value) and downregulated proteins (along the negative component value). Component two likely represents transiently regulated protein, which may peak at the 30 minute time point. The components tend separate from the 0 min control point along the negative values, and progressively return. Interestingly the 240 time point over shoots the 0 min control, and begins to separate in the positive direction along component two. This behavior may be reflective of more complex regulation. (D) The remaining components explain ~ 14 % of the variance are not readily interpretable, and although may contain relevant biological information, are likely due to experimental noise.

Two primary components, comprising 49% and 34% of the variance, respectively, were found (plotted with respect to one another in Figure 4.35, C). Consistent with the hierarchical clustering, component one separated the stressed time points from the control in a time dependent manner (later time points we separated to a larger extent). Component two separated the time points in an additional dimension, which appeared to reflect transient expression. Along component two, the 30 minute time point was maximally separated from the 0 minute control (at a value ~-100), and each successive point up to the 120 minutes time point began to fall closer to the value of the control (value ~ 50). Of distinction, the 240 minute time point actually surpassed the control along component two (falling at a value ~150). This phenomenon may indicate that component two comprises more complex behavior than simple transiently expression; for example, some upregulated proteins may actually overshoot their initial steady state value, and may be expressed at a lower magnitude once a new steady state position is reached. The remaining fraction of variance (only components three and four out are displayed in Figure 4.35D, 14~%) was not readily interpretable, and may represent a more subtle behavior or noise in the data.

As with the biological triplicate analysis of heat stress, the component loading values (PC1 and PC2) were plotted against one another (Figure 4.36) with the intention of visualizing outlying protein data points, which may be instrumental for explaining the biological process involved with each component. A number of heat stress regulated proteins were separated from the large cluster of unchanged proteins located at coordinates [0,0], and are highlighted on the plot. Many of these are those proteins which were also found to be regulated in the triplicate analysis of heat stress; these include proteins involved in catabolic processes and nutrient acquisition (TDH1, GPM2, HXT7, and ARO10), alternative carbon source utilization (BDH2, ACS1), NAD metabolism (BNA5), protein folding (HSP82 and SSA4), and arginine synthesis (ARG 3, 5 and 8), categories previously observed.

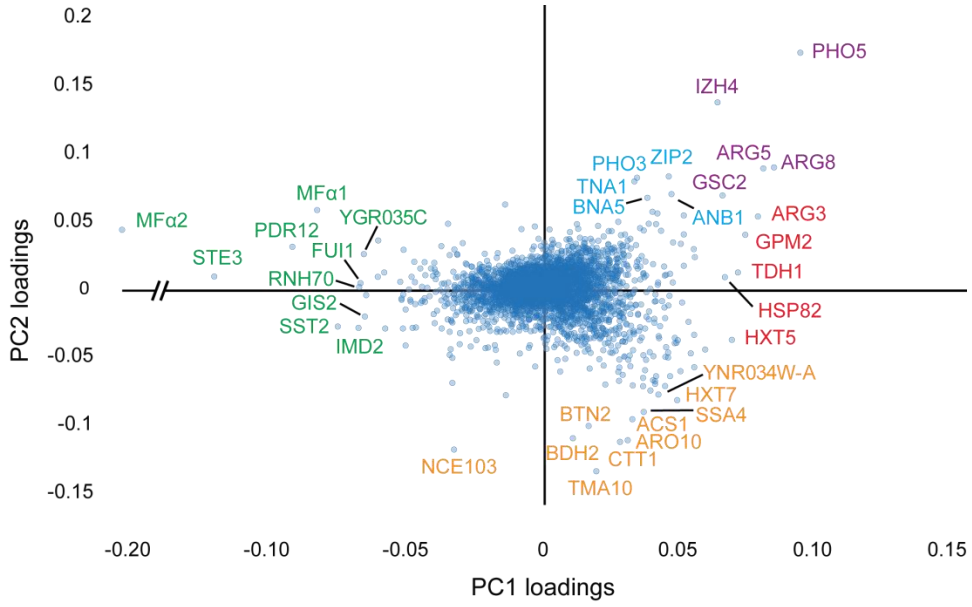


Figure 4.36. Principal component loading value plot of PC1 vs. PC2 loadings. The top 10 proteins from PC1 and PC2 are labeled (both positive and negative directions). Some proteins (PHO5, ARG8 and GSC2 for example) populate the top 10 list of both components. Color key: Red, contained within the top 10 loading values in only PC1 (positive direction); blue, contained within the top 10 loading values in only PC2 (positive direction); purple, contained within the top 10 loading values in PC1 and PC2 (positive directions); green, contained within the top 10 loading values in PC1 only (negative direction); orange, contained within the top 10 loading values in PC2 only (negative direction). Several of the highlighted proteins were also observed as significantly regulated proteins in the biological triplicate analysis of heat stress (MFα1, STE3, RNH70, HXT7, HSP82, SSA4, ARO10, ARG proteins, etc.). Some additional proteins were found to be regulated in the heat stress (generally at the later time points) including many involved in nutrient sensing such (PHO3 and PHO5), and additional proteins in the mating pathway (MFα2).

The same isoform-specific expression patterns discussed in Figure 4.30, such as the upregulation of TDH1 and GPM2 and not their counterparts (TDH 2 and 3, and GPM 1 and 3, respectively), were also observed here (Figure 4.37). Other proteins which were not quantified or not regulated within the time frame of the triplicate heat stress experiment, PHO5 and HXT5 (not previously observed) and PHO3 (not significantly upregulated until 240 min), were observed to be regulated in the time course. This observation of additional upregulated nutrient sensing proteins (PHO3 and PHO5) is consistent with the observation that stress mimics a nutrient deprived state. Finally, additional components of some previously observed pathways, for examples MFα2 in the mating pathway, were also observed here. This time course data set both confirms and complements the biological triplicate analysis of heat stress.

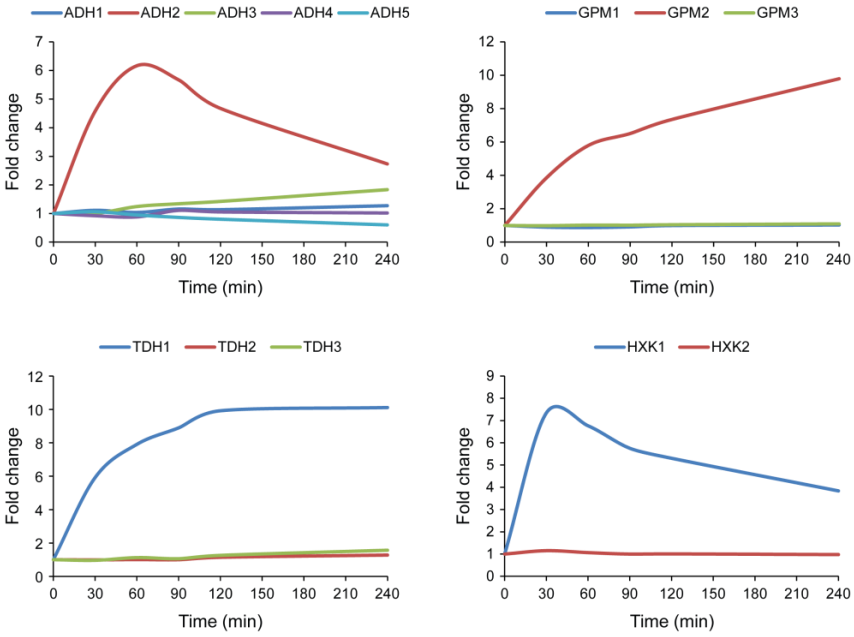


Figure 4.37. Examples of temporally regulated protein isoforms. These proteins isoforms were first found to be differentially regulated in the triplicate analysis of heat stress (Figure 4.30). In the same manner, specific isoforms of the exemplified proteins, all of which are involved in metabolic processes, are upregulated while others remain relatively unchanged. Some of the exemplified proteins demonstrate transient expression profiles (ADH2 and HXK1), while other are continuously upregulated (TDH1 and GPM2). These results confirm the previous observation of isoform specific regulation during heat stress.

Many interesting trends are observed through plotting the relative TMT intensities of proteins which comprise the top PCA loadings. As expected, the components reflect types of temporal regulation including continuous upregulation (PC 1 and positive directions, Figure 4.38, A and C), continuous downregulation (PC1, negative direction, Figure 4.38, B) and transient regulation (PC2, negative direction, Figure 4.38, D). There are examples of the proposed regulatory behavior (discussed above), including those upregulated proteins which fall below their initial steady state value once, presumably, a new steady state is reached. NCE103, a carbonic anhydrase exhibits such behavior (Figure 4.38, D), and is responsible for the hydration of CO_2 to bicarbonate, an important metabolic substrate for carboxylation reactions. NCE103 deletion strains are sensitive to H_2O_2 treatment, suggesting its role in the defense against reactive oxygen species[55], which are likely formed from the increase in metabolic activity during stress. NCE103 is undetectable under anaerobic conditions and poorly transcribed under conditions of high CO_2[55], presumably due to the increased formation of spontaneous bicarbonate at high CO_2 concentrations (downregulated by a feedback mechanism). Conversely, NCE103 is expressed under conditions of low CO_2[56]. The observed expression profile of NCE103 may be explained in the following

manner: NCE103 is upregulated after the initial temperature assault in preparation for or as a reaction to ROS, or alternatively in response to an increased need for bicarbonate. As the yeast continue to populate the media over time (increased density), and as a result of the increased respiratory activity (discussed in the triplicate analysis of heat stress), the CO_2 level may increase beyond initial conditions, resulting in the downregulation of NCE103 beyond its steady state expression level. Whether this behavior may be due to a biological necessity or biological/experimental stochasticity remains to be seen. It does, however warrant further investigation and experimental confirmation of the proposed NCE103-ROS-CO_2 relationship.

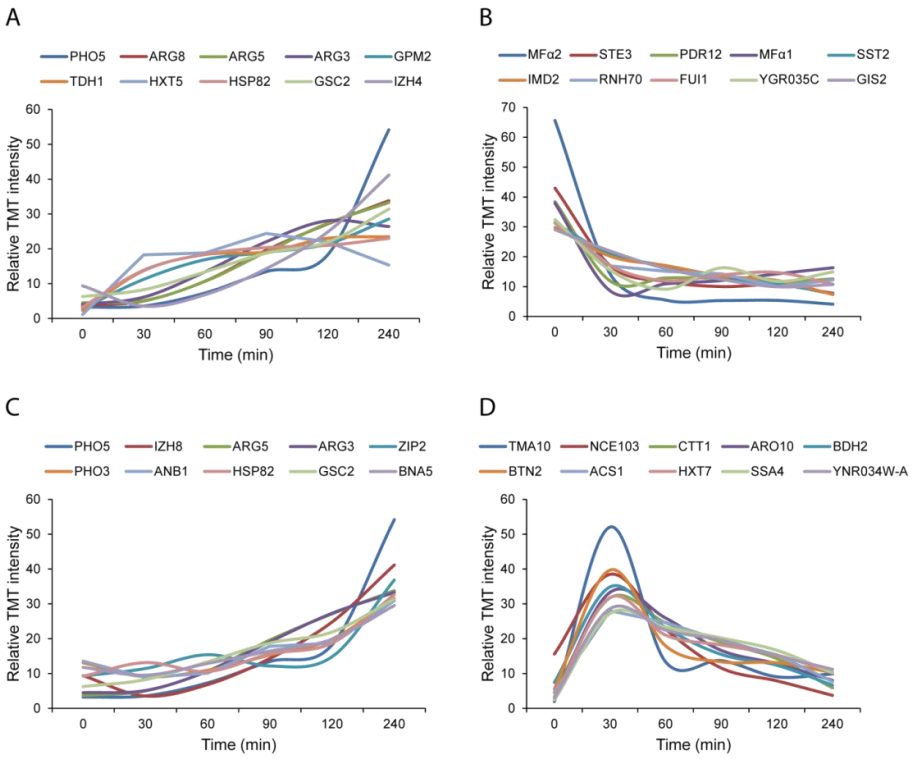


Figure 4.38. Normalized TMT intensity plots of the highlighted proteins reveal different modes of regulation associated with each component. (A)Proteins from PC1, positive direction, display a pattern of consistent upregulation throughout the time course. (B) Proteins from PC1, negative direction, display a pattern of downregulation, some which drop within 30 min and remain steady, and others which are gradually downregulated throughout the time course. (C) Proteins from PC2, positive direction, also display a pattern of upregulation over time, though there is spike of regulation between 120 and 240 minutes. This component seems to correlate with PC1 (positive direction), though is also consistent with discussed observation of protein magnitude changes thought the time course (Figure 4.34). (D) Proteins from PC2, negative direction, display transient behavior. In these expression profiles, the maximal expression occurs at 30 min, and many proteins return to steady state levels, or in some cases below, by 240 minutes.

Though interesting relationships were observed through PCA, selecting specific groups of proteins based on loading values, which represent each component can be difficult. Besides some of the obvious outliers highlighted in figures 4.36 -4.38, separating out relevant proteins from irrelevant proteins can be arbitrary, as subsequent PCA loading values often change by a small margin. In addition, the top ten proteins (based on loading values) from PC1 and PC2 (positive directions) displayed similar patterns, suggesting an additional analysis may be required. NMF was implemented as this complementary analysis to PCA.

The Use of Non-Negative Matrix Factorization (NMF) for the Analysis of Heat Stress Time Course Data

The objective of NMF, similar to PCA, is deconvolution of data through the reduction of many variables into a manageable number of components (often referred to here as a basis in NMF). The fundamental idea of NMF is that it decomposes a matrix (here the protein expression value matrix, using normalized TMT signal to noise) into two matrices under the constraint that the factorized matrices must contain all non-negative values. This procedure generally results in non-unique factorized matrices, which requires that the user pre-determine the rank, denoted k, of the decomposed matrices, which will result in k columns in one matrix and k rows in the other. In mathematical terms k defines the number of basis columns in one matrix, but k may also be interpreted as the number of clusters that NMF produces. One of the matrices produced by this process is called the basis matrix, and it has one row per protein and one column per cluster. The non-negativity constraint results in proteins being primarily upregulated in one cluster appearing with increased values in the basis matrix in the specific cluster they belong to. The other matrix is the coefficient matrix that is indicative of which cluster each original sample belongs to.

NMF is well suited for uncovering sets of proteins that are specifically upregulated in a given group of samples. In contrast to methods like PCA, the sets of distinguishing features for a group have

been found to be more readily interpretable[57]. For biological data the number of distinguishing features in a cluster can be reduced to those that particularly distinguish that cluster using one of a number of scoring functions[11]. This ability provides a certain amount of transparency to the algorithm's output, and allows more ready downstream analysis by standard methods like GO category enrichment.

A particular challenge in applying NMF is choosing the appropriate value for k. Because the algorithm generates non-unique solutions, it is run many times using random starting values (to find global vs. local minima, here 200 runs) and an average consensus of these solutions is used to estimate the parameters and group memberships. This process is repeated for several values of k and the consensus (Figure 4.40) and various other readouts for cluster stability (Figure 4.41) can be used to estimate an optimal number for clustering. In this manner, the challenge of NMF is also a benefit, in that contains metric for the discovering the appropriate number of clusters contained within a data set. Either the average solution or the solution with the least error (figures 4.40 and 4.39, respectively) can be used as the final clustering.

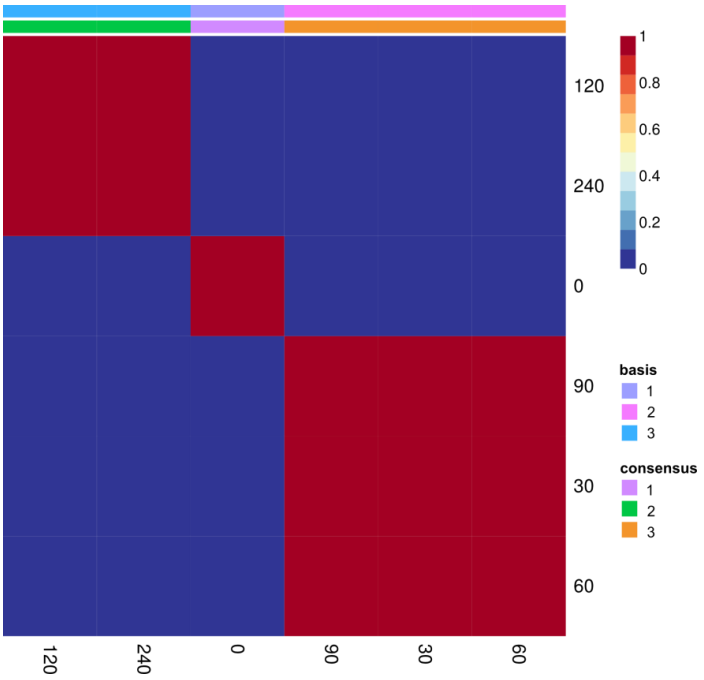


Figure 4.39. Best clustering consensus among NMF iterations. The heat map of the best connectivity matrix, indicating basis number and consensus groups, is displayed. The color values are based on which consensus the samples fall within. The map indicates that three consistent groups exist within the data, each of which likely represents a temporal mode of protein expression. The basis and coefficient matrices associated with this consensus map were used in further operations, such as feature extraction for identifying proteins which represent each group

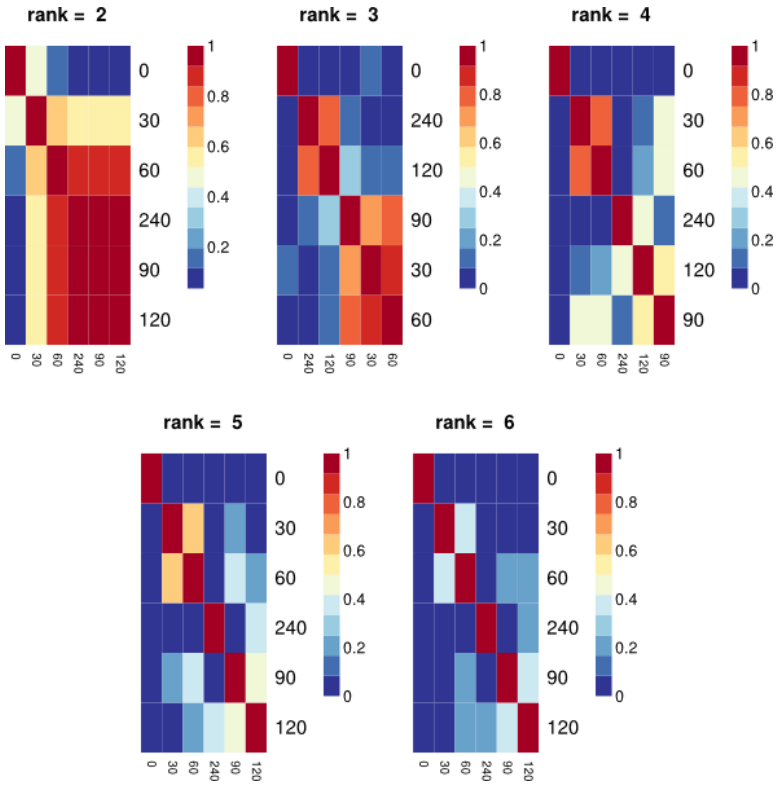


Figure 4.40. Reordered consensus maps of the time course data, using different numbers of clusters (K=2 to K=6). The average consensus map over all NMF iterations is presented for each number, k. The color bar represents the frequency in which two samples cluster together throughout the iterations. A rank of K = 3 showed the most consistent clustering, suggesting 3 groups of proteins exist in the data set, and is consistent with the best consensus map found among all NMF iterations.

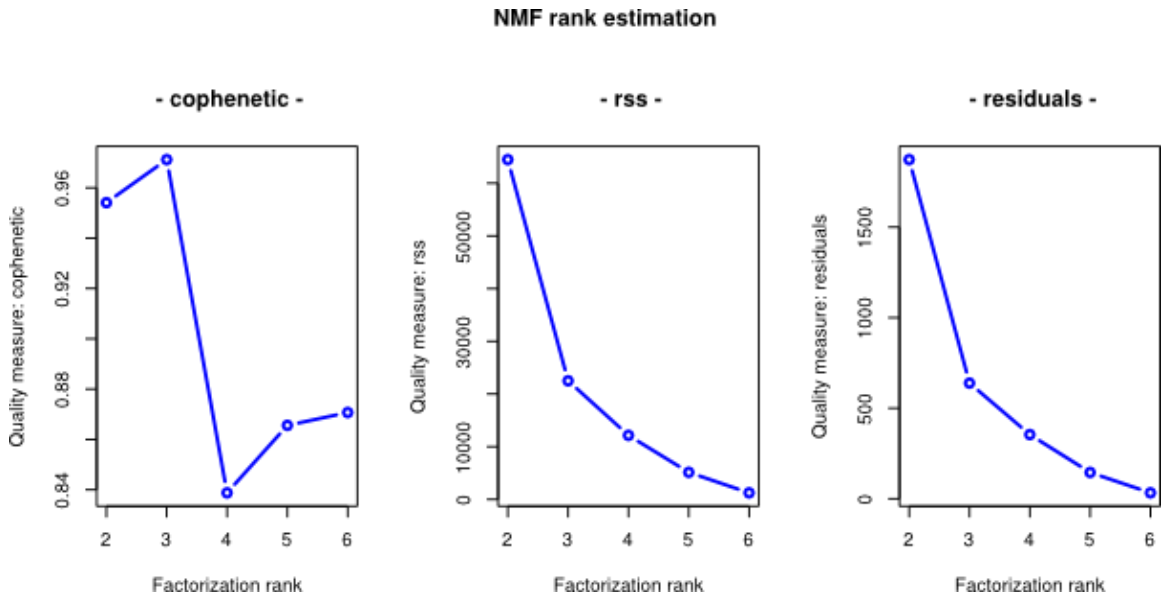


Figure 4.41. NFM rank estimation for the heat stress time course data, based on cophenetic correlation and the residual sum of squares/residual values. Ideally, the cophenetic correlation should be maximized, whereas the residual sum of squares (rss), or the residuals themselves should be minimized. However, to avoid over-fitting the data, the inflection point within the rss or residuals is often used. Both metrics support the use of three clusters for NMF, as the maximum cophenetic value and the rss/residual inflection point occurs at this number. The use of three clusters also fits the perceived biology of an upregulated, a downregulated and a transiently regulated category. Other means of evaluating clustering performance, such as dispersion, gave identical results.

For the time course data, it was found that clustering into three groups appeared to give the least error, (figures 4.39-4.41). The 0 min time point made up its own consensus group, whereas the 120 min and 240 min points formed another consensus group, and the 30 min, 60 min and 90 min points formed the final consensus group. Using different numbers of cluster (2, 4-6), the time points were ordered in an altered manner, though showed much poorer clustering consistency (Figure 4.40). The cophenetic correlation score (Figure 4.41), a measure of clustering stability, was maximized at K=3, supporting the use of a basis number of three.

As described above, the coefficient matrix contains information about the relationship between biological samples and the clusters, including the stability with which a given sample contributes to a cluster (Figure 4.42). For the most part, the 0 min point (to some degree the 240 min point as well) contributed to cluster 1. The 30 min point in particular contributed to cluster 2, followed by (in decreasing order of contribution) the 60 min, 90 min, and 120 min points. The 240 min, followed by 120 min point, and to some degree 90 min point, contributed to cluster 3. The 120 minute point, which contributed to both cluster 2 and cluster 3, actually contributed more to cluster 3 (based on coefficient values) than cluster 2 and hence clustered with the 240 min time point in a consensus group.

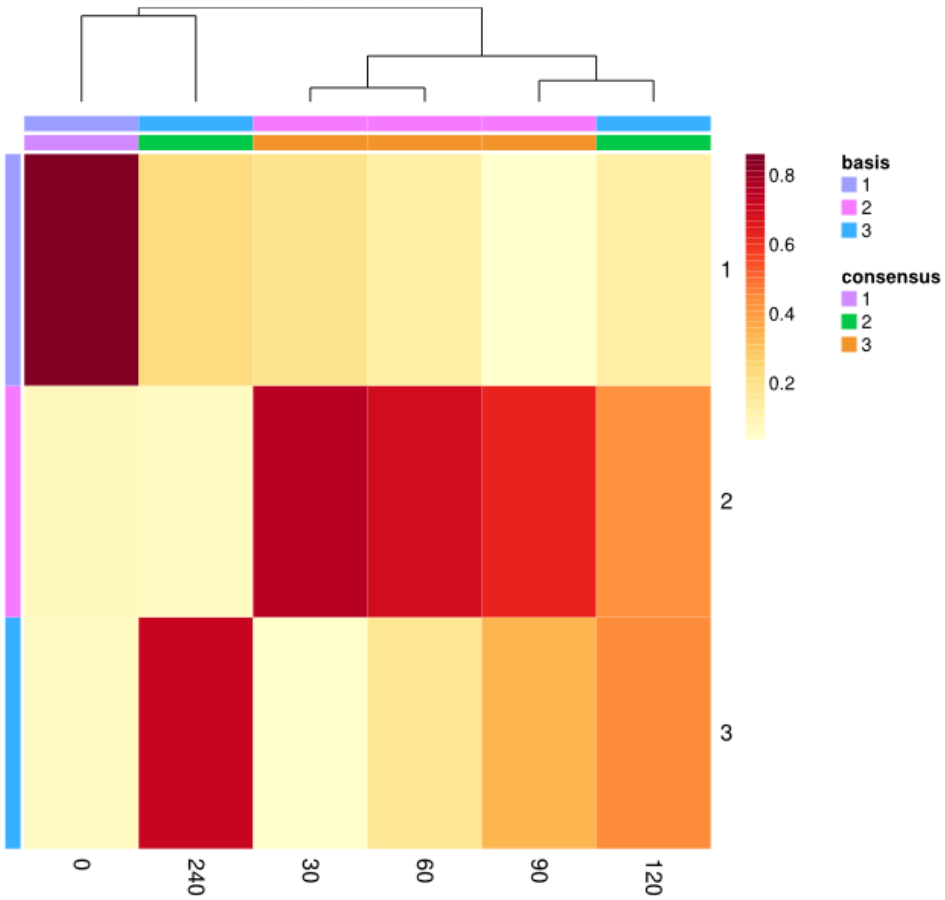


Figure 4.42. Heat stress time course NMF coefficient matrix. A hierarchical cluster of coefficient values is displayed on top of the diagram, as are the basis numbers (equal to the group number discussed later) and consensus groups. Numbers on the right side of the diagram also refer to basis. The 120 min point shares some coefficient values with the 90 min time point in basis 2, and thus the 120 min coefficient clusters with the 90 min point. The 240 min point shares some coefficient values with the 0 min point in basis 1, and thus clusters (albeit extremely weakly) with the 0 min point. In this case, the clustering and basis assignments are slightly different. Instead of simply following the clustering, the basis followed the consensus map. The 0 min time point was assigned to basis 1, the 240 and 120 min time points were assigned to basis 2 and the remaining points comprised basis 3. Though the 120 and 90 min point clustered with the 90 min point, based on a similar coefficient pattern, the 120 min time point contributed more to basis 3 than 2, and thus was assigned to basis 3 by NMF, as the analysis was constrained by three clusters.

From each basis, a group of representative proteins (often referred to as features) were extracted through the use of a pre-defined scoring algorithm[11], which identifies outlying values. The group number is equal to the NMF basis number. Plotting the relative expression (relative TMT intensities) of the proteins from each group reveals the type of temporal regulation represented within each group (Figure 4.43). Group one contains proteins which are downregulated, whose maximal and

minimal expression are observed at the 0 min and 240 min, respectively (N = 162). In an opposite fashion, group 3 contains proteins which are upregulated, whose minimal and maximal expressions are observed at the 0 min and 240 min, respectively (N = 133). Group 2 is the largest group (N = 258), and generally contains proteins that are transiently upregulated, and where maximal protein expression occurs at 30 min. Many of the proteins in this group return to their initial steady state expression levels by 240 min. Consistent with genomics data[23], transient protein expression may be the more important mode of regulation during the heat stress response. Groups of transiently downregulated proteins, however, were not observed (in any NMF iterations of K = 2-6), and thus many not be as important a mode of temporal regulation during the heat stress response (within the tested time frame). This observation contrasts the genomics data (in which transient downregulation was observed), suggesting a disconnect exists between transcript and protein levels (discussed later). The gene symbols for each protein in a group are summarized in Tables 4.3 and 4.4.

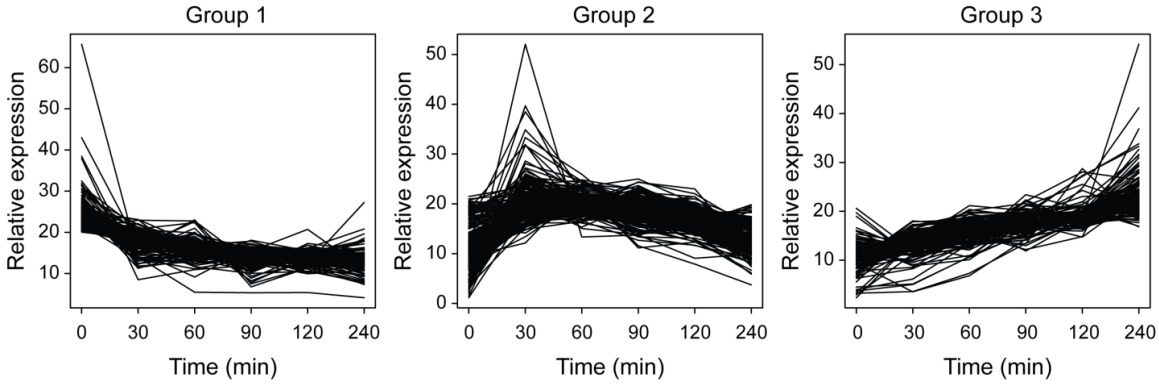


Figure 4.43. Expression profiles for extracted protein groups from basis 1, 2 and 3. The three groups fit a general profile of downregulated, transiently expressed and upregulated proteins, respectively. Group 2 contained the most proteins, and may possibly be subdivided into additional categories.

Table 4.3. Extracted groups of proteins from basis 1 and 3, which represent up and downregulated proteins. Features were extracted using a pre-defined score (see materials and methods). Many of the groups contents are consistent with previous observations, such as the upregulation of chaperones (HSP82, SSE2, SSA1), metabolic/nutrient components (GPM2, TDH1, PHO3, PHO5), redox proteins (TSA2, PRX1) and arginine synthesis proteins (ARG 5,8), all of which are contained within group 3 (N = 133). The downregulation of ribosome associated machinery (RPS and RPL, protein components of the ribosome, RMT2, NSA2, etc.) are contained within group 1 (N = 162).

Group 1

[1] MFα2 STE3 GIS2 SST2 IMD2
[6] ALT2 TRM2 AIM41 CPR2 YHB1
[11] KTR5 PDR12 NHP6A MFα RNH70
[16] FUI1 RRP8 HHO1 HOT13 COX5A
[21] RMT2 TOR1 GDH1 MRPL31 MNN4
[26] APA2 PUS9 NOP4 YGR035C NSA2
[31] YOR296W FYV4 SCW10 IPT1 TOS1
[36] REX4 PUS7 LEU1 ARN1 SNG1
[41] RPL40A SAM1 RPL37B SCP1 PFK26
[46] PDR15 EGT2 KEX2 TMA16 SWC3
[51] DBP2 NIS1 FTR1 HPT1 NAF1
[56] NSR1 YBL028C RPS26B FET5 POP8
[61] GAR1 CTS1 SAM3 BUD20 THI7
[66] IZH2 RPS29B TRP1 RSM27 MET10
[71] RPL16B OGG1 FET3 RPL34A KTR6
[76] SCW11 RIB4 RPL37A FEN1 SPO14
[81] MRP49 HMF1 SAH1 YER156C RPL10
[86] HHT1 EXG1 PLB2 SNQ2 RPL43B
[91] ALB1 NOP53 TIF4631 BAT1 RPL24A
[96] RPL16A SAM2 MDH2 FRE1 CBC2
[101] BDF2 YHC1 RSN1 TOD6 SIM1
[106] BAP2 HMT1 ATP14 SPE4 RAD27
[111] RPL6A MRT4 MTO1 HXT3 ITR2
[116] TYW1 GDT1 NOP2 CIC1 RPL35B
[121] OPT1 MAK16 RAX2 YCL019W MRPL10
[126] RKI1 MSS116 DBP3 NOP6 RPL14A
[131] AXL2 RLP7 HEM3 MSN2 RPS23A
[136] MRPL24 YGL101W UTR2 RPL34B LDB7
[141] BUL2 RPL19B KEM1 SAP190 OAC1
[146] SKO1 MAK3 CMD1 SCW4 RSM28
[151] NOP10 EPS1 NHP2 HEM12 DOT6
[156] RPS8A EBP2 URA1 CIN2 ACC1
[161] YDR026C ALA1

Group 3

[1] PHO5 ARG8 ARG5 IZH4 ARG3
[6] GPM2 TDH1 HSP82 GSC2 ZIP2
[11] PHO3 YRO2 TNA1 CPA1 AHA1
[16] ANB1 YGP1 SSE2 BNA5 SSA1
[21] HCH1 YPL067C YEF1 YNL054W-B YHR112C
[26] YIL055C MLC2 HPA3 CPR6 TSA2
[31] PRM8 PRX1 MAL32 HSP10 SCM4
[36] ALD2 YPL071C GRE3 LAP4 YBR016W
[41] BAT2 HSP60 ARG1 YER137C MRH1
[46] STI1 NTE1 BNA4 YCL047C OPI10
[51] CHA4 YNL200C FMP37 YDR098C-A HOR2
[56] HSC82 GPD1 AAC3 TPM2 SNZ3
[61] SED4 WTM1 CUP1-1 CRH1 BNA1
[66] YFR045W THI20 ARG82 PLB1 PTR2
[71] CTR1 CUE1 SPE2 STF1 YJR008W
[76] ARG2 YDR248C MTD1 YGL242C CTI6
[81] RNR3 HBN1 NAS2 RRD2 YGR169C-A
[86] MMS2 KAR2 PGM3 ADH3 AIM46
[91] SIS1 SIR4 YPL141C EXG2 SRY1
[96] DUG3 SMF3 DIA1 PYK2 RNR4
[101] YMR147W TAM41 BUD16 YGR031W DMA2
[106] KIP2 BUD31 IFH1 PTP1 COX15
[111] EAF6 PBI2 GRX1 OYE2 MIC14
[116] UBC7 YCL057C-A TAH1 SIP1 UBC8
[121] TPM1 DAK1 YDR186C ATG18 GON7
[126] TIM12 CYB5 MIG2 USA1 DCS1
[131] YPR127W SSA2 MKK2

Table 4.4. Extracted groups of proteins from basis 2, transient expression. Features were extracted using a pre-defined score (see materials and methods). The majority of previously observed upregulated proteins are contained within group 2 (N = 258, ARO10, HXT7, SSA4, HXK1, etc.), indicating that the majority of the heat stress response may by transient, as previously proposed by Gasch et al. based on genomic evidence.

Group 2

[1] TMA10 ARO10 CTT1 HXT7 SSA4 HSP26 YNR034W-A GPH1 YMR196W HSP12
[11] ACS1 HXK1 BDH2 NCE103 ADH2 YLR108C HBT1 HXT5 TFS1 HXT2
[21] GND2 PGM2 ALD4 BTN2 RTC3 YER067W GDB1 GCY1 CHA1 SOL4
[31] HSP78 HSP42 AIM17 PHM7 TSL1 GAD1 NQM1 GLC3 NCS6 RTN2
[41] ALD3 GSY1 GLK1 ARO9 PHM8 MSC1 APJ1 DCS2 YKL100C GLO1
[51] THO2 SSA3 STF2 YHR162W HIS4 GCV1 YMR090W SER3 IRR1 YLR460C
[61] CAR2 BRR2 YML131W COX5B CTR9 HIR3 PNC1 DUR1 CAT2 IRC20
[71] SNZ1 KAR1 HSP104 YBR085C-A PRM5 YNL134C HEH2 GCV2 ATR1 SPE1
[81] AMS1 RSC9 YGR250C SRC1 SWR1 GTT3 MTR4 OM45 BDH1 NOP7
[91] ITC1 UTP22 STH1 YMR291W PUT3 URB2 IOC2 ESF1 RAT1 HAP1
[101] ORC1 URB1 POL5 CPA2 YDR222W ORC5 SCC2 STU1 CBF2 HOS2
[111] HAL9 UTP20 CFT1 YKL151C COS9 PRP8 CIT2 RRP5 SYF1 CWP1
[121] MAG1 YDR391C CFT2 RPO21 POM34 NOC3 YNL022C STB4 NUP84 YCG1
[131] PIN3 NTH1 FMP52 OAF1 ALD6 KRE33 ADH5 TBF1 NOP14 GSH1
[141] FES1 HPR1 ISW2 CTF4 INP52 MSH6 RPA190 MSH2 ERO1 NOG2
[151] MOT1 SPG5 NOC2 ARE2 VID22 ADE17 STV1 GOR1 HSH155 TOP2
[161] GSY2 YGL140C RSC30 SEN1 DHR2 ASK1 HIR2 ECM16 TPS1 DBP9
[171] PAP1 POM152 CIN8 EMI2 INO80 CYC7 RSC3 NUP170 PUF6 SPT6
[181] RPB2 FUS3 NUP133 CIT1 RPC82 TPS2 ASN2 RRP12 IES1 RNY1
[191] PNS1 RKM3 GRE2 ISW1 YLR143W YOR052C SDA1 SIN4 RGR1 ASI1
[201] MDN1 YTA7 PMS1 MAK21 NCL1 UTP10 SPB1 SNF2 BRE2 NMA111
[211] YLR177W PRB1 NOP56 SIR3 SDS24 YLR278C APC5 DIP2 SET1 NUP120
[221] NUP157 YRM1 OM14 HDA1 CDC16 DEP1 UTP8 APC1 PTA1 BLM10
[231] MAK5 HMG2 PRP5 SPT16 PRD1 UGA1 CUP5 YJR015W MRE11 NUP85
[241] HDA3 YDL025C UGP1 NOG1 PDR1 SHM2 TRA1 PIN2 HST1 NOP9
[251] ULA1 RIX7 RPO31 RIX1 YLR446W SPP382 YFR039C ERG1

Expected proteins were found in each group, such as chaperones and catabolic enzymes in the upregulated and transiently regulated groups, and ribosomal machinery in the downregulated group. Additional gene ontology categories were extracted from these groups, compared to those already observed in the triplicate analysis of heat stress. From the group of downregulated proteins (group 1), for example, the biological process of pseudouridine synthesis (modification of uridine in RNA) was enriched 12 fold enriched. This category contains such proteins as NOP10, GAR1, NAF1, NHP2, PUS9 and PUS7**.** Pseudouridine is the primary modified nucleoside found in tRNA, and the downregulation of this process is consistent with an overall downregulation of translation. From the group of upregulated proteins (group 3), arginine biosynthesis was enriched 17 fold. Though arginine biosynthesis was discussed previously, the GO term itself was just below the $P < 0.05$ cutoff in the triplicate heat stress analysis**.** A variety of new GO terms were contained within the transiently regulated group of proteins, including transcription (2 fold enrichment), RNA elongation (3.5 fold enrichment) and chromatin remodeling complexes (3.5 fold enrichment) were found. Considering many of the previously observed metabolic proteins are also found in the transient group of proteins (categories for glycogen and trehalose biosynthesis were 11 and 9 fold enriched, respectively), it seemed odd that transcription processes were also present (as these categories were not picked up in the triplicate analysis of heat stress). To further understand this behavior, and to dissect additional patterns of temporal regulation, the proteins contained within group 2 were hierarchically clustered (Figure 4.44).

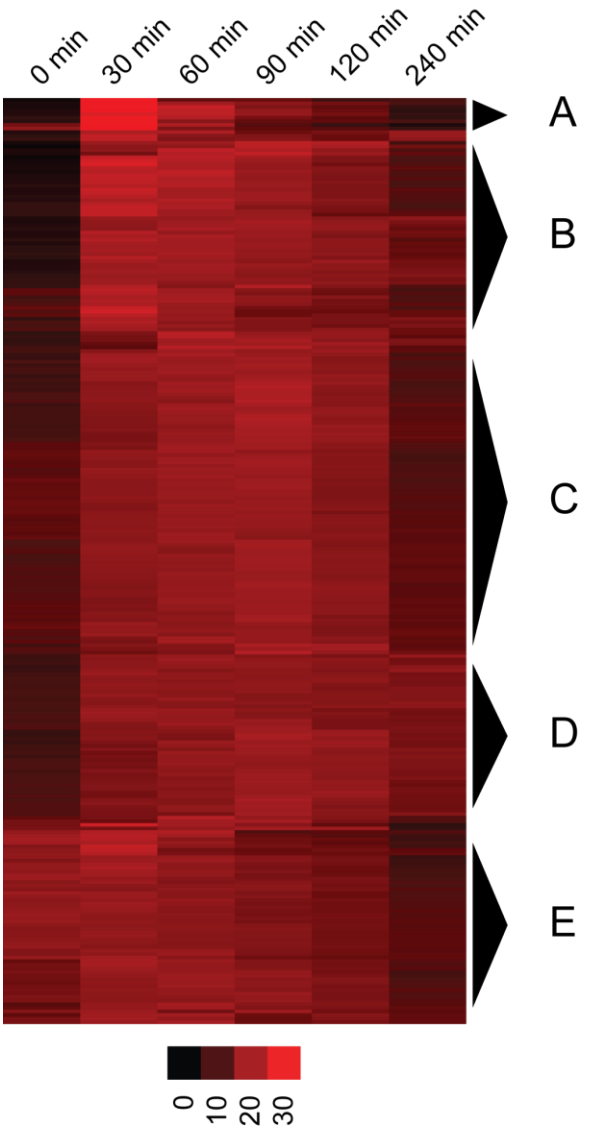


Figure 4.44. Clustering of transiently regulated group of proteins (NMF group 2) reveals additional levels of temporal regulation. The color bar represents % of total TMT signal in a given channel/time point. (A) Extremely induced transient expression, ~10 fold, e.g. SSA4 and ARO10. (B) Highly induced transient expression, ~3-6 fold, e.g. ALD1 and GSY1. The majority of the heat stress response previously observed are contained within these two clusters, protein folding, catabolic processes, etc. (C) Moderately induced transient expression, ~1.5-2.5 fold, e.g. PUT3 and THO2. The new categories involved with transcription, RNA elongation and chromatin origination are contained within this cluster. Though translation is generally shut off during heat stress (based on the down regulation of ribosome machinery), it seems likely transcription is not; conversely, certain aspects of transcription and chromatin remodeling may be positively regulated during the heat stress response, such as those genes involved in proline utilization, based on the upregulation of the PUT3 transcription factor. (D) Upregulated and sustained expression, ~2-3 fold, e.g. TPS1 and NTH1. Additional previously observed GO processes, such as those involved in glycogen and trehalose metabolism are present in this cluster. (E) Delayed downregulation, ~2 fold, e.g. NOG2, NOP7. RNA and Ribosome-associate processes are contained in this category.

From the clusters of the transiently regulated proteins, it is clear that additional categories of regulation exist: Extremely induced (~10 fold, Figure 4.44, A), highly induced (~3-6 fold, Figure 4.44, B) and moderately induced (~1.5-2.5 fold, Figure 4.44, C) transient expression, sustained upregulation (2-3 fold, Figure 4.44, D), and delayed downregulation (~2 fold, Figure 4.44, E). As discussed in Figure 4.44, many of these clusters contain proteins involved in previously observed biological processes, such as upregulated chaperones and catabolic enzymes, and downregulated ribosomal machinery. Within each cluster are functionally related proteins, such as the sustained upregulation (Figure 4.44, D) of TPS1, a

trehalose synthase component, and NTH1, the neutral trehalase (demonstrating the aforementioned paradoxical relationship between biosynthesis and utilization). Interestingly, the cluster of moderate transiently expressed proteins (Figure 4.44, C), was enriched with proteins involved in transcription and chromatin remodeling. Therefore, although translation on a whole may be downregulated, it appears transcription is not; It is likely that the transcription of certain genes, such as the upregulation of the proline utilization transcription factor PUT3 (found in the cluster), contribute to heat stress resistance; indeed this example is consistent with amino acid catabolic process also observed. Highlighted hits from Figure 4.44, demonstrating these additional levels of regulation are plotted in Figure 4.45.

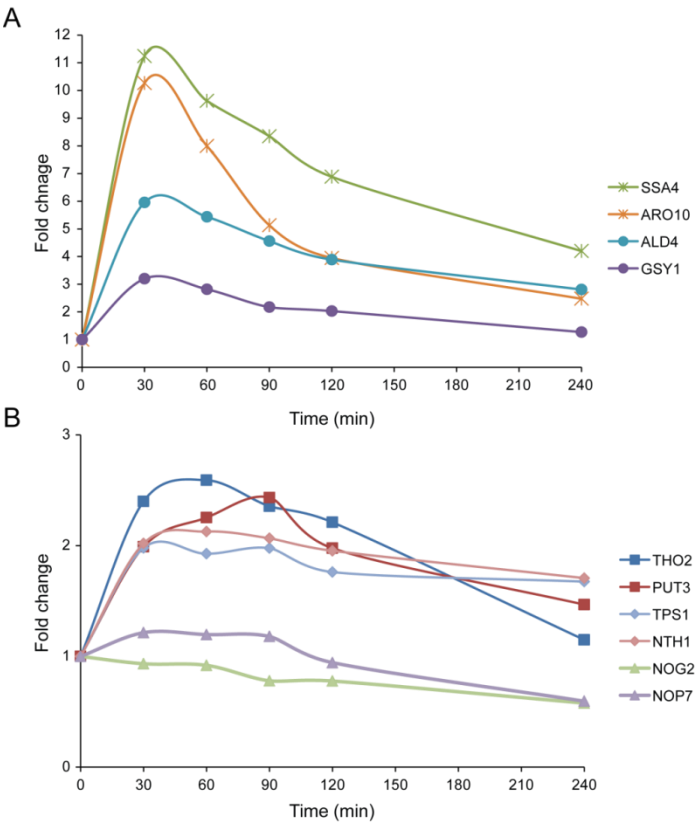


Figure 4.45. Plots of exemplar proteins from different transient (NMF group 2) protein clusters, demonstrate additional levels of temporal regulation. (A) Extremely and highly induced transiently regulated proteins. These proteins tend to be maximally expressed at 30 min. SSA4, Hsp70 family of molecular chaperone; ARO10, involved in amino acid catabolism; ALD4, alcohol dehydrogenase; GSY1, glycogen synthase. (B) Moderate induced transient, sustained and delayed changes. Proteins are upregulated by 30 min and maximum protein expression of moderately induced transient protein occurred at 30-90 min. Sustained changes were stable after 30 min. Delayed downregulated changes did not occur generally until after 90 min and minimum expression occurred at 240 min. THO2, transcriptional elongation; PUT3, transcription factor involved in proline utilization as a nitrogen source; TPS1 trehalose synthase (trehalose biosynthesis); NTH1, trehalase (trehalose utilization). NOG2, GTPase involved in ribosome export from the nucleus; NOP7, involved in ribosome large subunit maturation.

Although it may seem odd that the last two categories are contained within the transient expression group (NMF group 2), their specific expression profiles reveal the reason for their inclusion. Group 1 and 3 from the NMF analysis contain proteins which are steadily down and upregulated,

respectively, throughout the time course. The sustained upregulated proteins and delayed downregulated proteins do not display such behavior, but rather show characteristics (as far as NMF is concerned) of transient behavior; namely an obvious point in time where their expression changes (30 min for upregulated and 120-240 min for downregulated proteins). It is possible that if a replicate of the time course was included, or more proteins from the dataset displayed the sustained/delayed regulation, an NMF analysis using a basis number of 5 would have given the most consistent clustering. Thus these groups would have been directly identified from that analysis, simplifying the process. This last point highlights the increasing need for replicates in large scale biology.

Comparison of Publically Available Genomics Data with Acquired Proteomics Data

As suggested earlier, there is potential and indeed biological precedent for differential regulation between transcription and translation. To evaluate the extent to which this statement is true in the heat stress response, the data collected in this experiment was compared to publically available genomics data (Gasch data set[23]). In the Gasch heat stress time course 0, 5, 15, 30 and 60 minute points were acquired. 3, 638 transcripts and proteins were shared between the two data sets. The 30 min genomic and 60 min proteomic points showed the greatest correlation. The correlation, however was unequal between upregulated transcripts/proteins and downregulated transcripts/proteins. The upregulated transcripts (those which changed by at least 2 fold) were moderately correlated with protein upregulation (Figure 4.46, A). Downregulated transcripts and proteins, however, showed no overall correlation. It is reasonable to theorize that although protein upregulation is an active process by which new molecules must be created, protein downregulation (as a whole) may be a passive process; moreover, the downregulation of a protein, which requires its degradation, may not be actively controlled in all cases. Thus this difference in genomic and proteomic downregulation may be explained by protein half-life, after transcription and translation ceased. It would be of great use to the scientific

community to conduct a proteome wide survey of endogenous (untagged) protein half-lives in yeast, to confirm such a hypothesis. Alternatively, observed downregulation could be a result of the normalization scheme, as discussed in Figure 4.21 of the triplicate heat stress experiment, though the consistency of observed biology between experiments suggests otherwise.

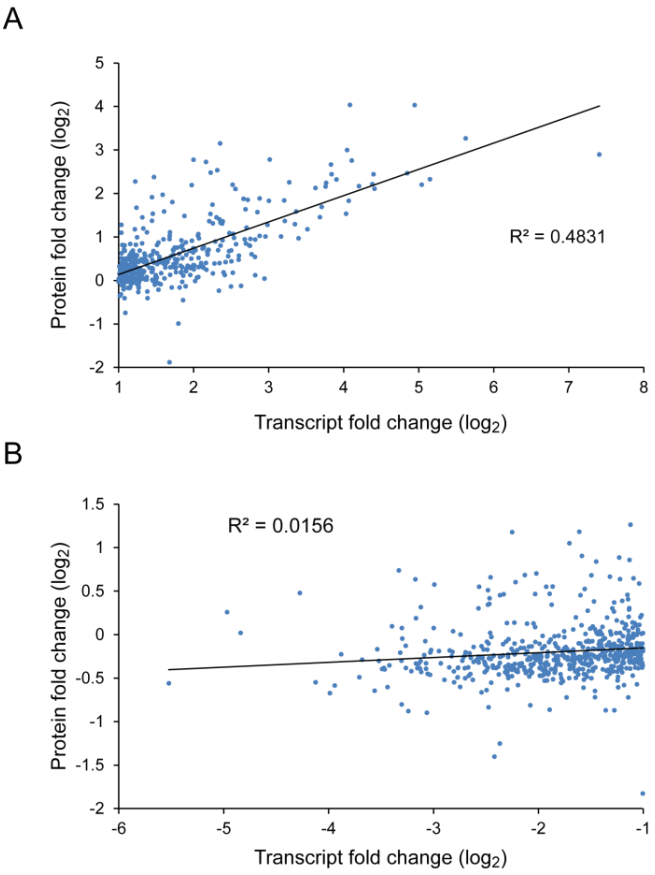


Figure 4.46. Correlation of upregulated and downregulated transcripts vs. proteins. Transcript data (publicly available) were plotted against the protein data from this analysis, and were filtered so that only proteins whose transcript changed by 2 fold or greater (log_2 +/- 1) remained. The 30 minute time point from the genomics analysis was plotted against the 60 min proteomic time point, as these samples had the greats overall correlation. (A) A moderate correlation exists between the transcript and protein levels in response to heat stress. (B) In contrast, generally there was no correlation between the downregulated transcripts and proteins, though in some examples the protein and gene are both downregulated. Once the transcription of a gene is ceased, the downregulation may occur through protein degradation, based upon a protein's half-life affecting this correlation between downregulated transcripts and proteins.

When the transcript and protein expression levels are clustered together, this discrepancy between the genetic and protein responses to heat stress is observed on a large scale (Figure 4.47). Although many of the extremely and highly induced proteins in heat stress clusters contain comparable transcript upregulation (Figure 4.47, A and B), clusters containing highly downregulated transcripts often did not contain downregulated proteins (Figure 4.47, D). Intriguingly, a cluster of moderately expressed proteins was identified which does not contain associated transcript regulation, indicating the possibility for posttranscriptional regulation of protein (e.g. mRNA or protein stability). To further understand the

differences and similarity between transcript and protein level regulation, the expression data from the protein groups identified through NMF was clustered with the respective transcript data.

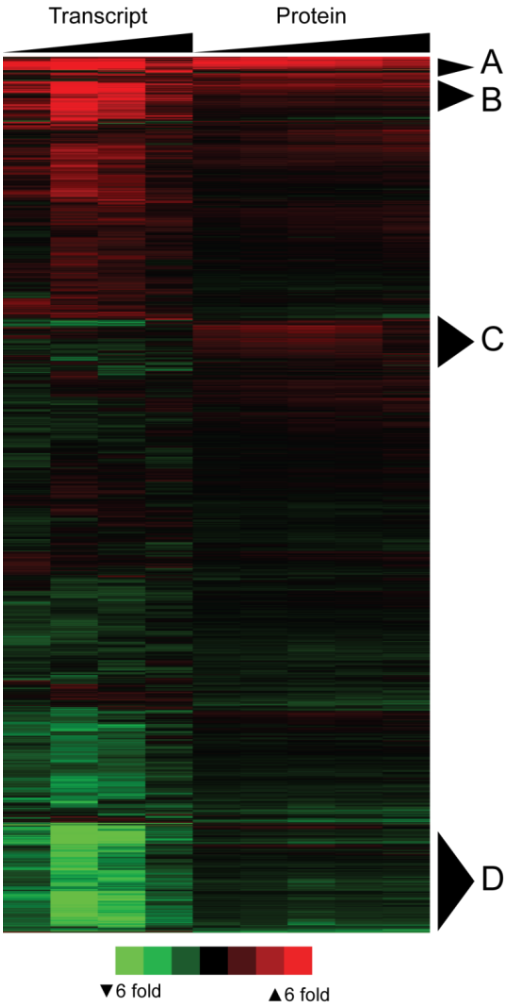


Figure 4.47. Hierarchical clustering of transcript and protein responses to heat stress demonstrates little overall correlation between protein and transcript regulation. Samples are ordered by increasing time. All values are $\log_2$ ratios of each time point compared to the unstressed control. Although ~1000 transcripts changed by 2 fold or greater at least one time point, only several hundred proteins demonstrated similar regulation. (A) Extreme transcript and protein induction. Proteins in this cluster were upregulated (at maximal expression) by greater than 4 fold. (B) Extreme transcript and large protein induction. Proteins in this cluster were upregulated by 2-4 fold. Transcripts levels generally were induced by a greater magnitude than protein levels, usually 8 fold or more. (C) Uninduced transcripts and induced proteins. Proteins in this clustered were upregulated 1.5-2.5 fold. (D) Greatly repressed transcripts and generally unaltered protein expression. Though some examples correlate well, this cluster shows little correlation between transcript repression and protein downregulation.

Despite the lack of correlation between downregulated transcripts and proteins on a whole, some proteins did follow their mRNA counterparts, albeit by much smaller magnitudes (Figure 4.48, A). The upstream transcripts of proteins which demonstrated a delayed downregulation pattern (Figure 4.44, E) are also downregulated. These downregulated proteins which are correlated with repressed transcripts are involved in the plethora of ribosomal processes discussed. There was no constancy among protein half-lives in this cluster, however, using the one available large-scale study[58]. Either the proposed connection between protein downregulation and protein half-life is incorrect, or the result

demonstrates the need for new measurements of protein half-lives using endogenously encoded proteins (as TAP tagged proteins were used in the cited analysis). Alternatively, proteins in cluster A may be actively degraded (altered half-life) during the early heat stress response, whereas those in cluster E may be degraded later. If this behavior is true, it may support the discussed hypothesis that some ribosomal proteins compete for molecular chaperones during the stress response, and their active removal is a reflection of limiting this competition.

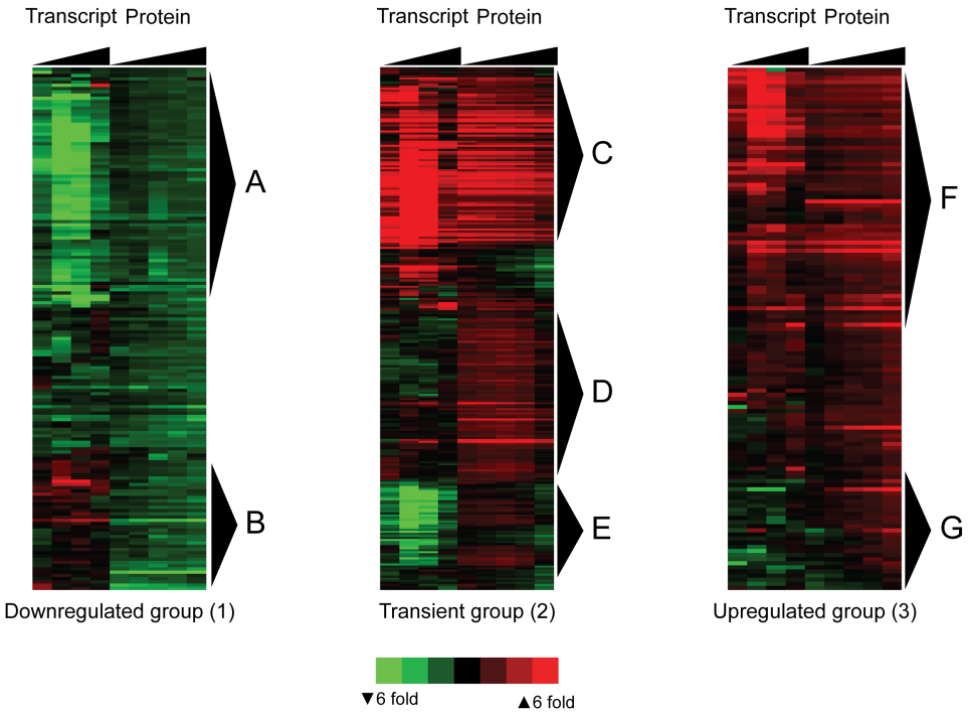


Figure 4.48. Clustering of protein expression data from different NMF groups with the respective transcripts data reveals specific discrepancies between gene and protein level responses during heat stress. Samples are ordered by increasing time. (A) Both transcripts and proteins are downregulated in some cases. The genes/proteins in this cluster are those involved with the many aspects ribosome assembly, maturation, export and function discussed. (B) Protein downregulation without associated transcript changes. Proteins in the mating pathways (MFα 1 and MFα 2, STE 3), and interestingly TOR1 are in this cluster. (C) Upregulated transcripts and proteins. Both transcripts and proteins are transiently expressed. The majority of the heat stress response is contained within this group. (D) Uninduced transcripts and upregulated proteins. The proteins involved with transcription, RNA elongation and chromatin remodeling are contained within this cluster. (E) Transiently downregulated transcripts and delayed downregulation of protein expression. Additional ribosomal components are included in this cluster. (F) An additional cluster of upregulated transcripts and proteins. Though transcripts still demonstrate transient expression, proteins in this group are consistently and steadily upregulated over time, and included the majority of NMF group 3 proteins. (G) An additional cluster of uninduced transcripts and upregulated proteins. Such proteins as PHO3, PHO5, and ARG 8 are contained in this cluster.

Other proteins which may be subject to regulation through degradation are those downregulated proteins whose transcripts are not repressed (Figure 4.48, B). Proteins in the mating pathway are contained within this cluster, and their relevance in not readily understood. TOR1, however is also contained within this group, and may be indicative of interesting biology. Heat stress has been shown to phenocopy rapamycin treatment[42], and rapamycin treatment induces many of the same protein which are induced by heat stress[43]. Knockouts of TOR1 and SCH9 (yeast homolog of S6K and AKT) have been shown to increase life span and heat stress resistance[59]. Thus, a connection exists between the two pathways, and may help explain the nutrient deprived state which heat stress mimics. The regulation of TOR at a posttranscriptional level may allow yeast to respond to nutrient conditions more rapidly, than through transcriptional regulation alone.

In an opposite fashion, groups of upregulated proteins with no upstream transcript changes were found (Figure 4.48, D and G). The majority of the proteins contained within these cluster are those associated with transcription, RNA elongation and chromatin remodeling. The significance of this behavior is unclear and warrants further investigation. It may simply reflect a mechanism to more rapidly upregulate transcription machinery and by consequence their downstream targets. Though transcripts were generally induced in a transient manner, upregulated proteins in this analysis (downstream of these transcripts) demonstrate both transient and consistent upregulation (Figure 4.48, C and F). The cluster of transiently expressed transcripts and proteins were more highly upregulated. Both clusters contain the wealth of the stress proteins already discussed, though the transiently expressed protein cluster contains more transcripts/proteins. Specific examples of transcript vs. protein regulations are highlighted in Figure 4.49. In general, this analysis of transcript and protein levels comparisons highlights the need for the simultaneous acquisition of genomics and proteomics data sets. An analysis of directly paired samples would be of great value.

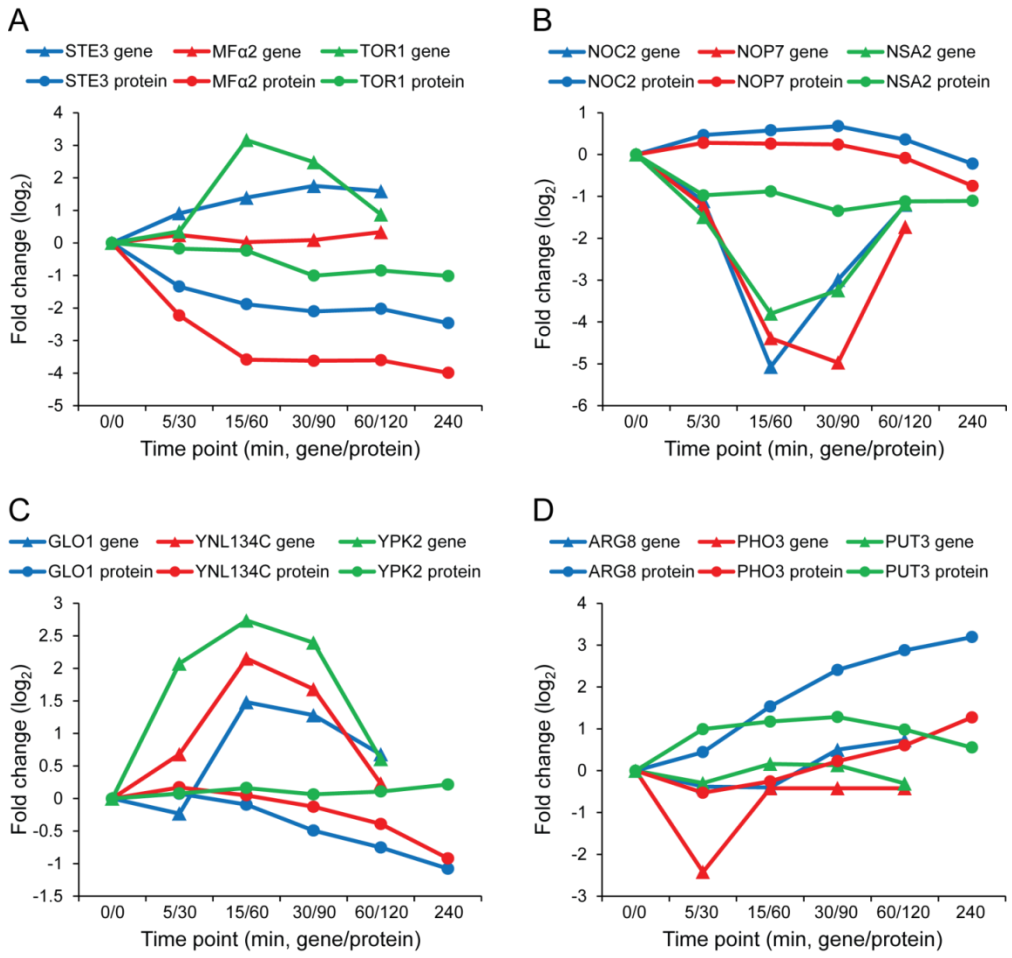


Figure 4.49. Examples of differences between transcript and protein regulation during heat stress. (A) Protein downregulation without upstream transcript downregulation. STE3 and MFα2 are in the mating pathway; TOR1 is involved in nutrient/growth pathways. In many cases transcripts were actually observed to be induced. (B) Downregulation of transcripts and proteins. NOC2, NOP7 and NSA2 function in ribosomal associated processes. Though the protein is downregulated, it is at a reduced magnitude compared to the upstream transcripts. (C) Transcripts induced without downstream protein upregulation. GLO1 is an osmotically regulated glyoxylase, and may be specific to osmotic stress; YNL134C, uncharacterized protein; YPK2, kinase involved in cell wall integrity. (D) Protein upregulation without upstream transcript induction. ARG8 is required for arginine biosynthesis; PHO3 is involved in the phosphate starvation pathway; PUT3 is a transcriptional activator of proline utilization genes. These examples highlight the potential for posttranscriptional regulation, through translation efficiency or protein turnover, for example.

A Proteomic Analysis of Multiple Stress Conditions Reveals Common and Unique Stress Responses

The final demonstration of TMT for proteome wide multiplexing involves the comparison of cold, oxidative (refed to as "H_2O_2 stress"), osmotic (refed to as "salt stress"), heat, and cytotoxic/ER stress (referred to as "canavanine stress") to an unstressed control. The intention of this experiment is

not to dissect one condition at a time systematically, but rather to investigate commonalities and differences amongst the conditions. Taken in context with the deeper analyses of heat stress, the ability exists to assess which fraction of that response is specific to heat, and which may be a more general program of stress adaptation. This experiment also provides a large proteomic screen of stress that may inspire hypotheses and future experimentation.

In all stress conditions tested, the aforementioned trend of a greater number of upregulated proteins compared to downregulated proteins held true (Figure 4.50). As suggested, this feature may be due to the active upregulation and passive downregulation of proteins as a whole (notwithstanding genomic changes, or an effect of the normalization scheme). In contrast to the other stresses, cold stress did not exhibit a wide array of protein changes, and had a narrow distribution of protein ratios; it is likely that the wrong time point (one hour) was selected for the cold stress response, and a cold stress time course may be useful as a future analysis. Canavanine stress demonstrated the widest distribution of protein ratios, followed by heat, salt and H_2O_2 stresses. This behavior is likely a reflection of the stress responses themselves and not experimental error, as quantification error as a whole should usually be equal amongst all TMT channels. That being said, unforeseen biological effects (e.g. particular sensitivity of canavanine treated yeast to handling) could contribute to these differences. Inclusion of replicate experiments would confirm if this notion is correct, reiterating the increased relevance of replicates in large-scale biology.

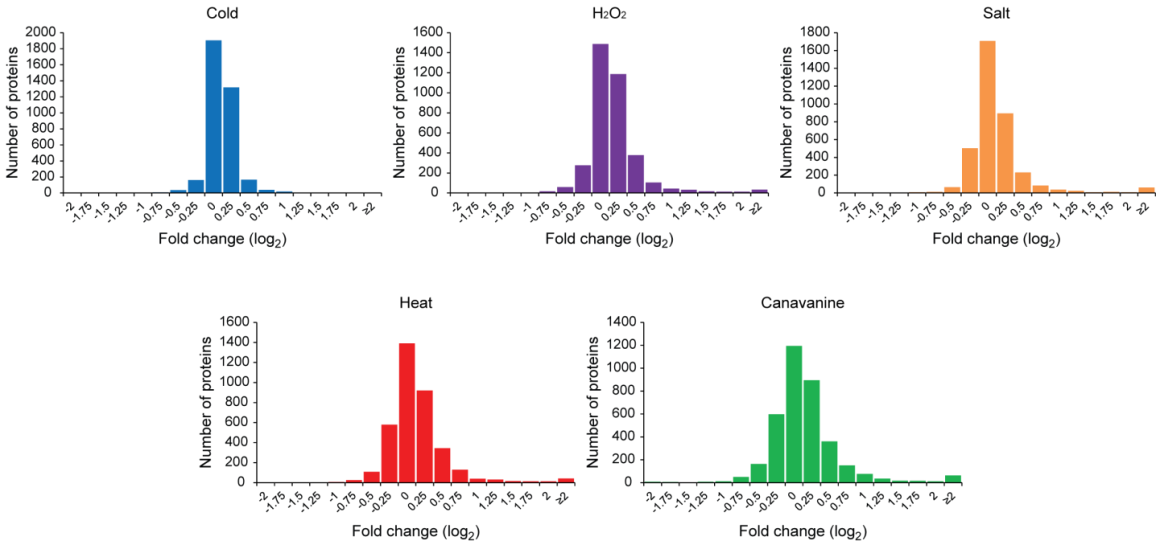


Figure 4.50. Distribution of quantified proteins from the 5 stress data. As previously observed, more significantly upregulated proteins are contained in the data set than down regulated proteins (2 S.D. for significance). This feature is now well established in the stress response, and is likely reflective of the requirements during a transient stress event prior to reaching a new steady state. Few proteins were regulated in the cold stress response within 1 hr, suggesting it may be a slower process. Canavanine stress caused the most significant change, showing the widest distribution of protein ratios. Heat and salt stresses had a similar distribution of protein ratios, followed by H_2O_2 stress, all of which were intermediate to canavanine and cold stress. In general, it is clear that the majority of the proteome is unaffected by stress.

In accordance with the histograms of protein ratios, canavanine stress induced the most relevant (2 S.D.) protein changes, whereas H_2O_2, salt and heat stresses contained a similar number of regulated proteins, with respect to one another (Figure 4.51, A). In all conditions, though, over 100 proteins changed by relevant magnitudes. Proteins regulated by canavanine stress tended to change by a greater magnitude compared to the other stress conditions. These data may indicate that canavanine is particularly toxic to yeast, and response to its presence requires greater protein regulation. Many of these stress regulated proteins were unique to one condition (Figure 4.51, B), which may suggest they are required for the response to specific environmental stimuli. In contrast, many proteins were also regulated in 4 conditions, presumably H_2O_2, salt, heat and canavanine stresses, suggesting these proteins may be general stress response proteins. Additionally, a large number of proteins were also

regulated in two stress conditions; likely, these proteins are those regulated in similar stresses, such as heat and canavanine stresses.

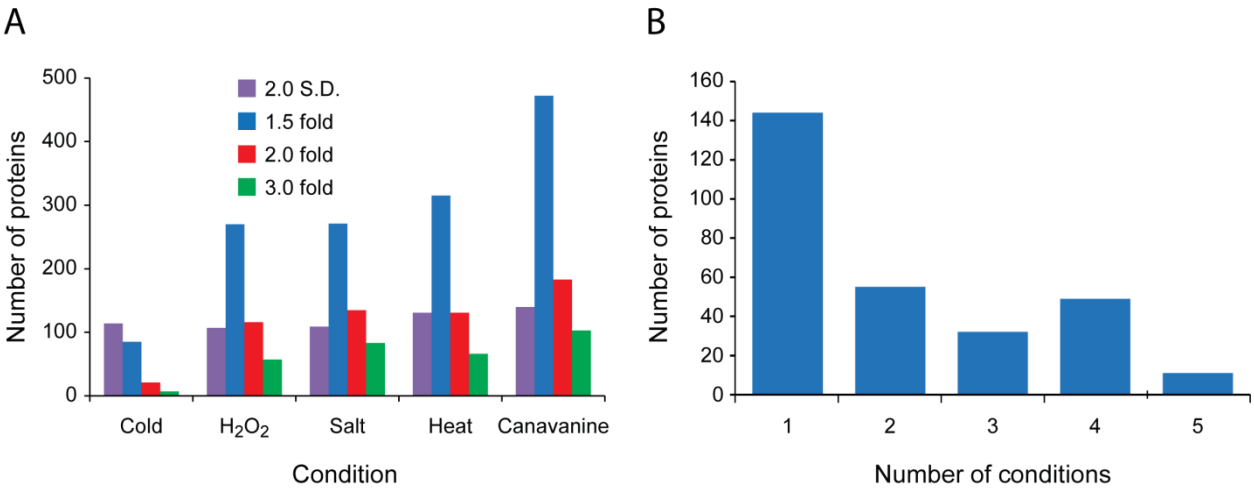


Figure 4.51. Data set statistics of regulated proteins identified in the yeast five stress experiment. (A) The fewest significantly (+/- 2 S.D) regulated proteins were observed in cold stress. The most significantly regulated proteins were observed in the canavanine stress condition. H_2O_2, salt and heat stress contained a similar number of significantly regulated proteins. Following this trend, the magnitude of protein change was highest in canavanine stress (those regulated by the indicated fold change), lowest in cold stress (very few proteins changed by greater than 1.5 fold), and similarly intermediate in the remaining stresses. (B) Many of the significantly regulated proteins were unique to one condition, suggesting these may be particularly useful for adaptation to a given environmental stress. There was also a large number of significantly regulated proteins common to 4 stresses conditions, presumably H_2O_2, salt, heat and canavanine. These are likely general stress response proteins. Proteins which were significantly regulated in two stress conditions may be those involved in adaptation to heat and canavanine stresses, as these are likely the most related amongst the stress conditions tested.

Though analyzing proteins on an individual basis may reveal relevant results, it is difficult to assign the relative contribution a protein may make to the stress response on its fold change alone. By analyzing the data using PCA and NMF, it is possible to characterize which biological processes are general stress responses, and which may be more specific to a given stress state.

A Comparison of Two Yeast Stress Data Sets Supports Their Combined Use in Downstream Analyses

When initially attempting to interpret the stress data thorough PCA, the results were unremarkable. Without replicates or related samples (such as time points in the heat stress time course), it was difficult to understand what fraction of the variance was due to true biology, and what

fraction may be a results of either experimental noise or biological stochasticity. Fortunately another stress data set was obtained (using a 2 hour time point) in parallel to the discussed data set, which showed similar enough expression profiles to be used for comparison. Only proteins quantified in both experiments (N = 3, 445) were included. Though there are differences in protein expression, the relative TMT intensities in a given condition showed a high degree of correlation between the two stress experiments (Figure 4.52).

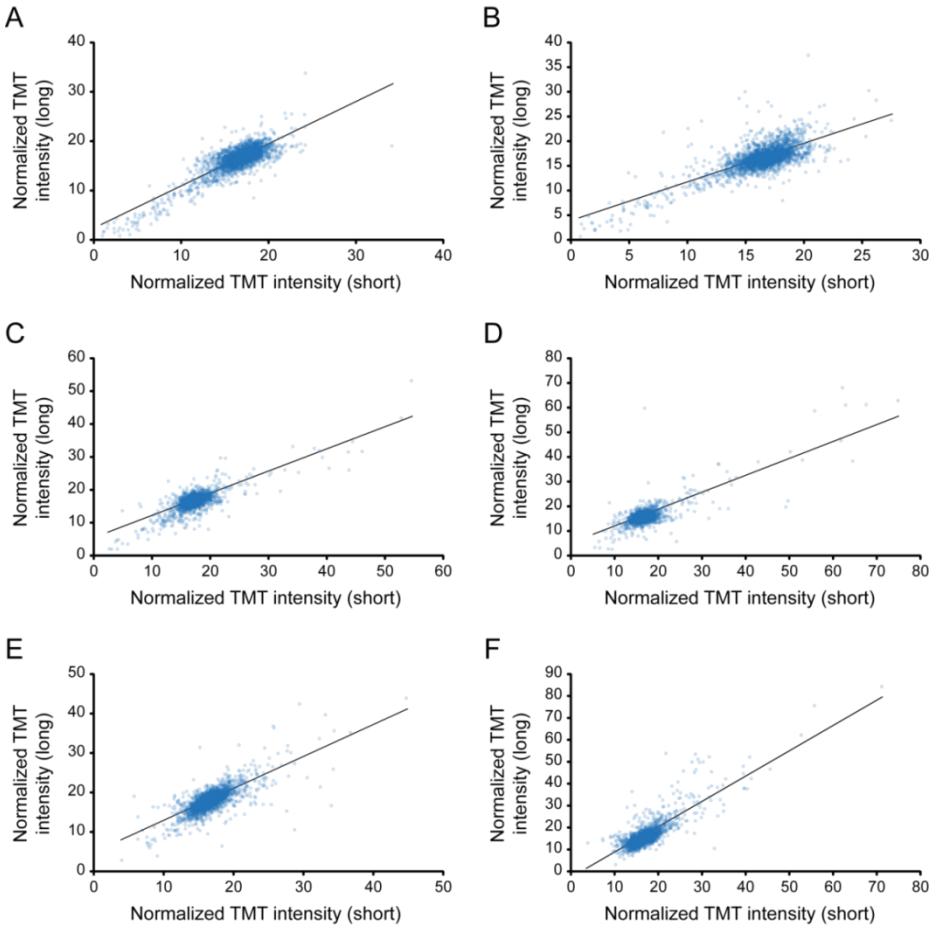


Figure 4.52. Correlation between short (1hr) and long (2 hr) stress points using normalized TMT intensity. (A) Control, (B) cold, (C) H_2O_2, (D) salt, (E) heat and (F) canavanine stress replicates are compared. Linear regression produced a correlation value (R^2) of ~0.6 among the stresses, suggesting a high degree of correlation, though some scatter does still exist. The distribution of values suggests a few outlying proteins may be adversely affecting the correlation, as the majority of the data clusters on the regression line. Proteins quantification based on single peptide measurements were omitted to avoid stochastic variance, though many single peptide protein quantification were still consistent among replicate. The regression line slopes and intercepts were ~0.8-1.2 and ~-2-4, respectively.

Residual plots demonstrate that the majority of data cluster along their respective linear regression lines (Figure 4.53 residual ~ 0), and that the trends are skewed by a few outliers at very low, and occasionally very high relative TMT intensity (as judged by the moving average of the residuals); often these cases represent border-line quantification, those with lower signal to noise quantification in many of the channels. Very few values fall outside one standard deviation (red lines in Figure 4.53).

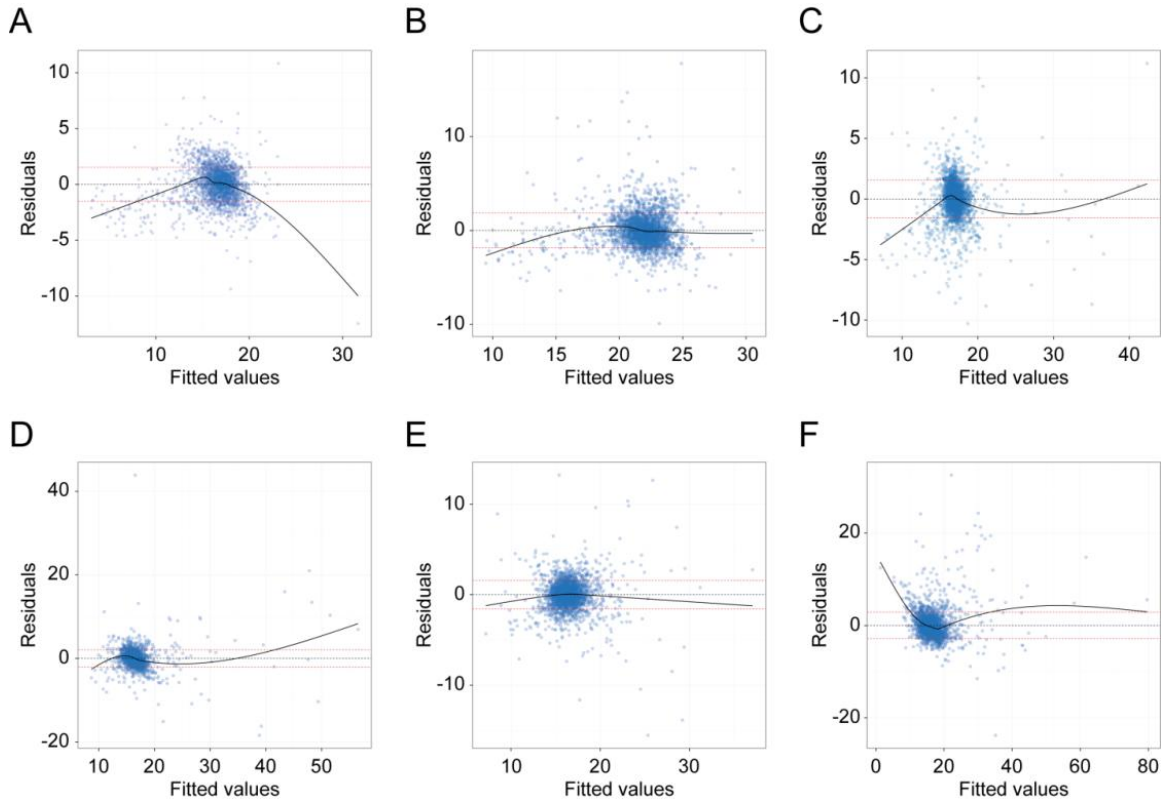


Figure 4.53. Residuals values from the linear regression analysis of stress replicates. Residual values are the distance each point in Figure 4.52 falls from its linear regression line. Fitted values are the normalized S/N from previous plots. Residuals for the (A) control, (B) cold, (C) H_2O_2, (D) salt, (E) heat and (F) canavanine stresses are displayed. The values are centered at a residual value of 0 (dashed black line). Dashed red lines represent one standard deviation from that line. Most of the data is fixated around the 0 axis and within one standard deviation, thus demonstrating the reproducibility between short and long stress experiments. The solid black line is a moving average of the residual value. Often outliers at very low and very high fitted values adversely affect the average residual value, and likely represent poor quantification events.

Recall that a 100 S/N summed TMT S/N cutoff was implemented to filter out most poor quantification events. Occasionally, however, a peptide will pass this cutoff with poor signal in the majority of the channels, but sufficient signal still in other channels to obtain a 100 S/N summed value.

As poor signal in one channel affects the relative signal of another channel, the inclusion of these peptides adversely affects the reproducibility of a protein's quantification between experiments; it is evident, however, that they are infrequent. It may be of value to identify data set dependent TMT signal filter in the future (perhaps specific to each channel), or perhaps establish a TMT quantification false discovery rate. Such a method for estimating the quantitative false discovery rate may be based on metrics such as peptide to peptide TMT signal to noise variance, and could be trained against a data set of known TMT ratios. One metric being explored is the cosine distances between peptide TMT vectors in six dimensional space.

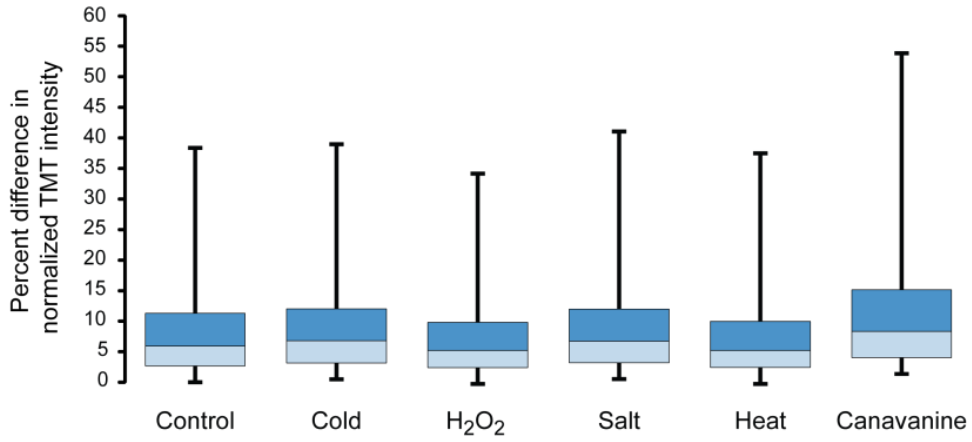


Figure 4.54. Box plots of normalized TMT intensity difference among short (1hr) and long (2hr) stress points. Maximum and minimum values (error bars) values were capped at the 98th and 2nd percentile, respectively. This consideration was required to accurately represent the spread of the data by removing the few outlying values (> 100% difference between experiments, due to poor quality data). The median values tend to fall at a ~5-7% difference between experiments. Even at the 75th percentile, differences <10-15% are observed between experiments. These distributions support the idea that the majority of data between experiments is reproducible between all stress states (including the control), and a few outlying values negatively affect correlation between experiments.

On a whole, the distribution of the differences in relative TMT intensity (for a given condition) between experiments follows the same trend as the residuals (box plots, Figure 4.54). The vast majority of proteins agree in their relative expression under all conditions between experiments (median differences, ~5-7%, 25th and 75th percentiles, ~2-5% and 10-15%,respectively, depending on condition), with the exception of a few outliers. Additionally when stress/control ratios are analyzed for their

correlation between experiments (Figure 4.55), they produce a greater correlation than that of the relative TMT intensity plots (Figure 4.52). The reason for this correlation increase is due to the decoupling of control and the chosen stress condition (heat and salt are shown in Figure 4.55 for example) TMT intensities from the other stress conditions; moreover, variance from one channel is propagated to other channels when the relative values are used. It is, however, beneficial to use normalized TMT intensities for many analyses, such as hierarchical clustering, so that the control channel values are weighted in such methods.

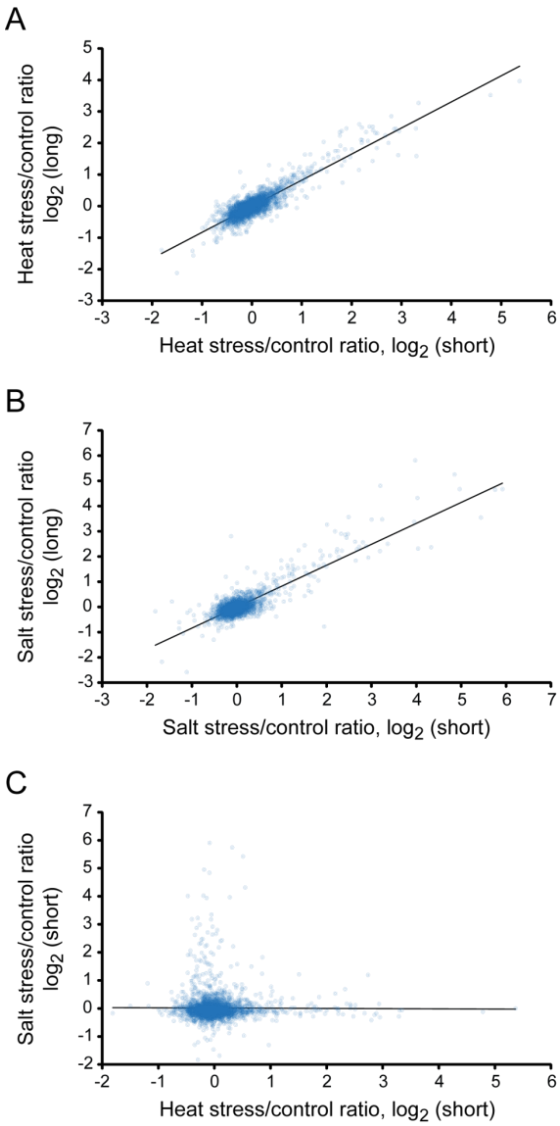


Figure 4.55. Stress/control ratios display a higher degree of correlation between short (1hr) and long (2hr) stress experiments. Using ratios in place of the normalized TMT intensity to compare reproducibility between experiments increase the correlation coefficient (R^2) to between 0.7 and 0.8, depending on the exact stress state. This increase in correlation is attributed to the decoupling the values of one stress state from the others. When using normalized TMT intensities, poor signal in one channel affects the signal in the other channels. Thus the relative TMT intensity in the other channels may be increased, whereas the ratio in those channels is unaffected (as the normalized TMT intensity in the control is also increased by the same fraction). The majority of the time, however, the normalized TMT values are more useful for comparing between stress states. (A) Heat stress and (B) salt stress ratios correlate well between replicates. (C) Heat stress and salt stress ratios do not correlate with one another, demonstration the described behavior is not simply due to the use of ratios vs. normalized TMT intensities. Proteins quantification based on single peptide measurements were omitted to avoid stochastic variance, though many single peptide quantification were still consistent among replicate.

Regarding hierarchical clustering, the experiments strongly cluster by their condition (though cold stress clustered with the control) and not by experiment (as can be the case) further supporting their joint use in further analyses (obvious clusters are present in both replicates in a stress specific manner, Figure 4.56). These described comparisons justify the use of the second stress data set as a replicate for further analyses.

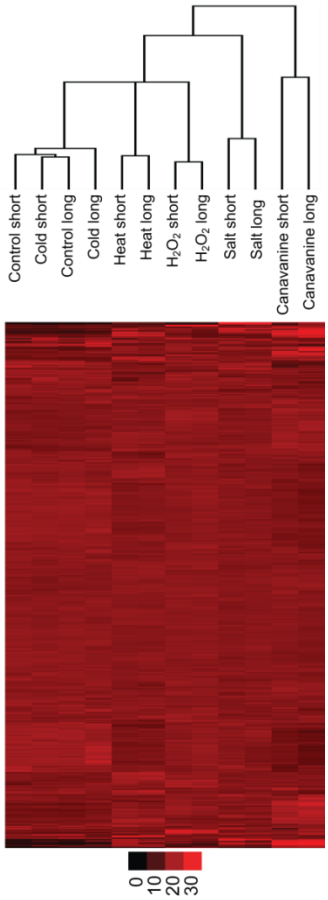


Figure 4.56. Hierarchical clustering of short (1hr) and long (2hr) stress data sets. In all cases, with the exception of cold stress, the correct stress replicates clustered with one another. All stresses (except cold) were well separated from the control. Interestingly the heat and H_2O_2 stress clustered together, and the salt and canavanine stresses were more divergent. It was expected that the canavanine and heat stresses would cluster most similarly. Indeed there is evidence in the heat map of similarities between heat and canavanine stresses, suggesting the similarities between the heat and H_2O_2 stresses may be more subtle. Only proteins which were commonly quantified in both experiments were included. Though this method of comparison is most correct, omitted values could affect the dendrogram.

Principal Component Analysis of the Stress States Reveals Unique and Shared Stress Responses

With the inclusion of both data sets, four primary principal components (out of a possible of 12) were characterized, which explained ~70 % of the variance within the two replicates of the five stress states (Figure 4.57, A). As expected, based on the small magnitude of protein changes observed, the cold stress condition often clustered with the control along the principal components.

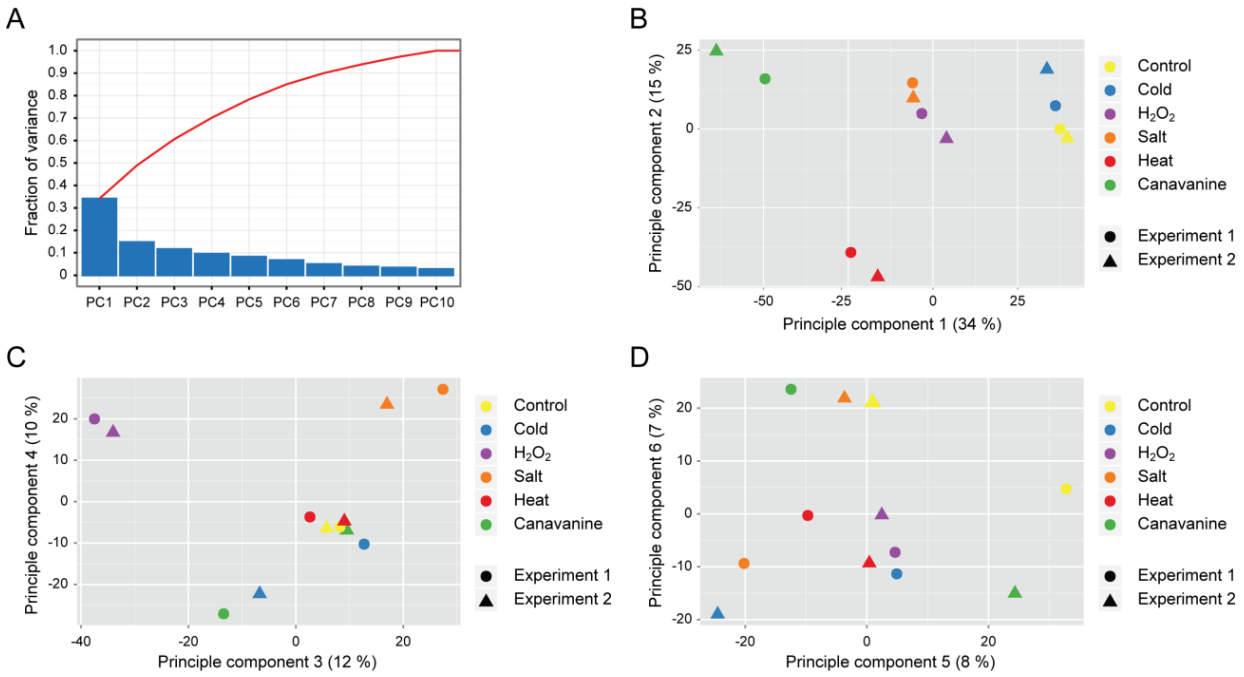


Figure 4.57. Principal component analysis the five stress experiment. The short and long stress experiments were treated as replicates for PCA, and are labeled experiment 1 and 2, respectively. In general each experiment fell within close proximity to one another on the component plots. The cold stresses did not separate from the controls, and are thus generally not further discussed as a stress responder. (A) The first four components have readily available biological interpretation and are responsible for ~70% of the variance within the data sets. (B) Component one likely represents a general stress response, as all stresses are separated from the controls. The canavanine stress was furthest separated by this component, followed by heat stress, and then H_2O_2 and salt stresses together. Component two separated out heat stress from the other stresses, and likely explains additional unique features in the heat stress response. (C) Component three separated H_2O_2 (negative direction) and salt stress (positive direction). Component four further separated salt and H_2O_2 stresses from the other conditions. (D) The remaining components are not readily interpretable, and may represent experimental noise or stochastic biological events.

The main component (PC1) explained the most variance (34 %), compared to other components, and separated the stress states (with the exception of cold) from the control. The stress states were not equally separated, however; canavanine was greatly separated from the other stresses, followed by heat stress, and salt and H_2O_2 stresses were together. As all stresses are separated along this component, it likely reflects a general stress response. If this is indeed true, it would suggest that the largest difference between adaptive states is not necessarily the variety of proteins which are regulated, but perhaps to what degree a common set of protein are regulated (protein abundance). That being said, the other components likely reflect some unique adaptations in each stress state. Component two,

responsible for 15% of the variance (Figure 4.57, B), further separated heat stress from the other states. Component three, responsible for 12% of the variance (Figure 4.57, C), primarily separated H_2O_2 from the other stresses (negative component values), though also separated salt stress (positive direction). Component four, responsible for 10% of the variance (Figure 4.57, C) also separated salt stress (primarily) and H_2O_2 stresses (to some degree). The remaining components (Figure 4.57, D) did not display readily interpretable behavior, and thus the remaining variance is unexplained and may be due to either noise in the data, or a stochastic aspect of biology.

Plots of the component loadings (Figure 4.58) did not reveal any obvious outlying proteins, which may suggest that the stress states are indeed highly related to one another. As a result of the nebulous distribution of component loading values, it was decided that the proteins comprising the top 100 loading values would be used for further analysis of the components, as often ~100 proteins changed in each stress condition. The use of NMF solves this problem of arbitrary feature selection and is discussed later.

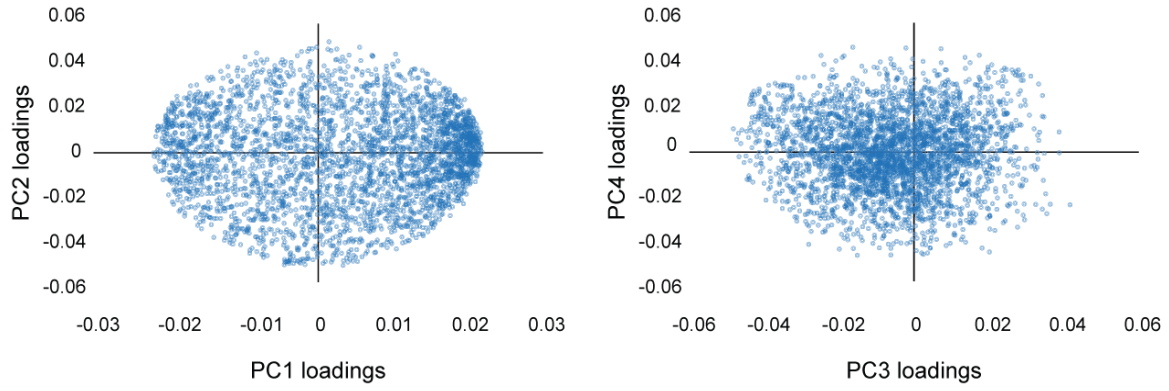


Figure 4.58. Principal component loading plots do not reveal obvious outlier points. As is often the case, the variance explained in principal component analysis of the five stress data set is more subtle, with many proteins contributing to the separation observed in each component. No single protein or small group of proteins is easily identifiable as the main proponent of the variance observed in a given principal component. This gradual effect creates more of a data cloud than a trend line. Thus it is difficult with solely PCA to determine the exact proteins responsible for separating out the stress types and to what degree each impacts that separation.

To ascertain the relationship among each stress state which is explained in component one, the distribution of normalized TMT values for the top 100 negative component loading values were plotted (Figure 4.59, A). Consistent with the idea that the magnitude of regulation among common stress proteins is primarily responsible for stress adaptation, all stress states demonstrate elevated expression levels of these proteins over the control, but to varying degrees. The relative expression among the stress states followed the component plot (Figure 4.57, B), where proteins were upregulated to the highest degree in canavanine stress, followed by heat stress, and salt and H_2O_2 stresses. The top 100 positive component loadings were plotted in a similar manner (Figure 4.59, B), demonstrating an equal but opposite trend. Likely these proteins represent general up- and downregulated stress proteins, respectively. To understand the biological roles of these proteins, GO analysis of the top 100 negative and positive loading was conducted.

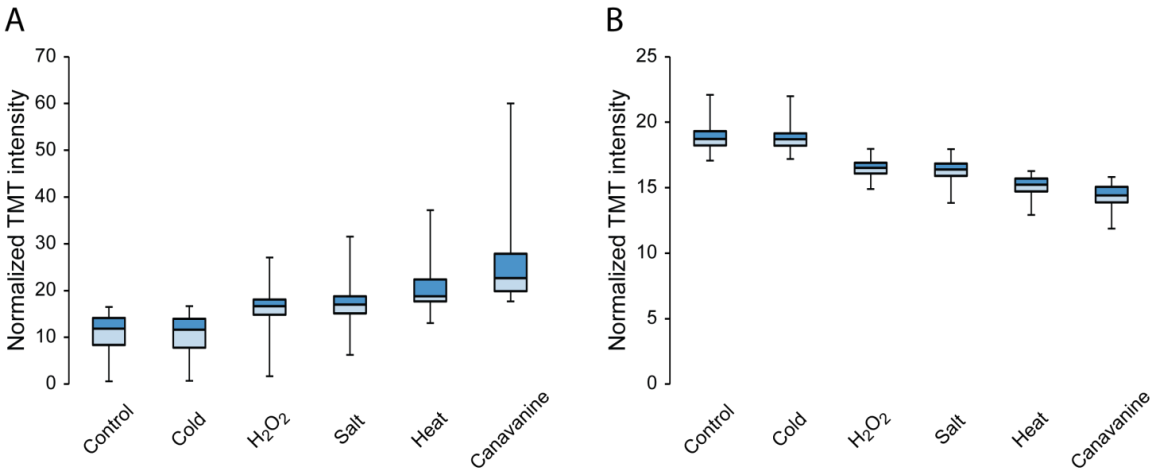


Figure 4.59. Box plots of normalized TMT intensities from the top 100 proteins in component one (based on positive and negative loading values) explain separation amongst most stress states from the control. The 25th, 50th and 75th percentiles are plotted by the box boundaries, and the error bars represent the minimum and maximum values. (A) Upregulated proteins (relative to the control, negative loading values in Figure 4.57) and (B) downregulated proteins are plotted (positive loading values). In component one, all stress states with the exception of cold stress (which generally showed little change) were separated from the control. In particular the canavanine treated samples were separated in this component, followed by heat stress, and salt and H_2O_2 stresses. The proteins responsible for the separation observed in component one are likely general stress response proteins, in which canavanine treated yeast displayed the greatest magnitude of change in both up and downregulated directions. The magnitude of change for down regulate proteins was lower than that of the upregulated proteins as previously observed.

Many of the GO categories observed in the analysis of common upregulated stress proteins (Figure 4.60) are those which where previous found in the heat stress response (response to heat ~6 fold enriched). Though canavanine contained the greatest magnitude of protein change, many of the proteins found in the GO analysis were regulated in all stress states (with the exception of cold, Figure 4.60, A).Thus, although the level of importance these proteins play in a given stress state differs, they all may be required to some degree. Catabolic Kegg pathways for starch/sugar metabolism were highly enriched (~8 fold). Other metabolic processes such as the use of alternative carbon sources (pentose catabolism ~25 fold, alcohol catabolism ~4 fold) were also present. These suggest that in all stress states, the mimicking of nutrient limitation discussed in the context of heat stress may be present. Interestingly, the categories of trehalose and glycogen biosynthesis were highly enriched (~18 fold and ~16 fold, respectively), demonstrating that paradoxical behavior between their biosynthesis and utilization may be a common mode of stress adaptation. The vacuolar lumen (~30 fold), as well as vacuolar catabolic processes (~ 10 fold) and autophagy (~ 4 fold), were enriched. Protein folding (~ 4 fold) was enriched as well. These categories suggest that in all stresses, some level of protein misfolding occurs, thus requiring chaperones for refolding and processes such as autophagy for dealing with aggregated proteins. Thus there may be a mechanism by which the general response to stress, such as increased ROS production from heightened metabolic activity, cause protein misfolding, though likely to a lesser extent than elevated heat or canavanine incorporation (based on chaperone expression levels in the stresses, e.g. SSA4 and HSP104).

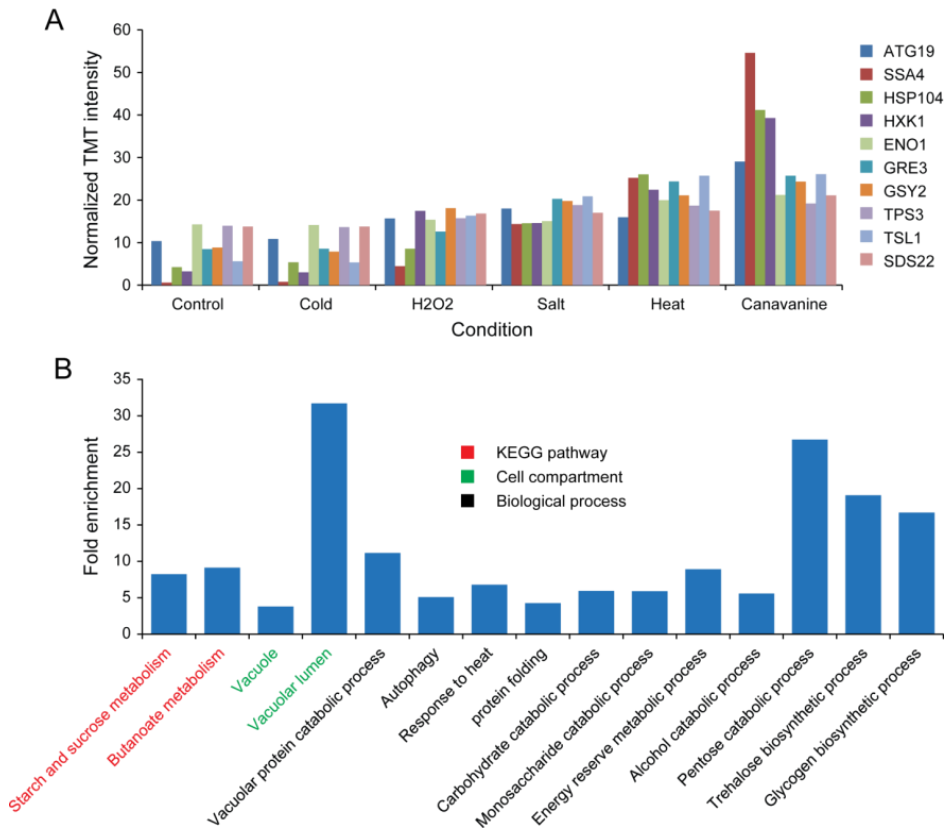


Figure 4.60. Gene ontology analysis of the general, upregulated stress response. (A) Highlighted hits from the top 100 principal component one loading values (negative direction), and (B) associated gene ontology. Many of the categories now seen as a general response were those observed previously in the heat stress experiment. These data may suggest what is traditionally annotated as heat stress, may in fact be a more general program of environmental stress adaptation.

Not surprisingly, many of the GO categories observed in the analysis of common downregulated stress proteins (Figure 4.61) are also those which where previous found in the heat stress response. These are the categories which pertain to the biosynthesis and regulation of ribosomes. Most of these categories were enriched between 5 and 10 fold. Ribosomal proteins themselves, rRNA processing enzymes, tRNA methylase, ribosome exporters, and many of the other discussed components are present. Interestingly, components of RNA polymerase I, responsible for the DNA depended transcription of rRNA, were generally downregulated. As rRNA is not translated, its transcription is a central point of regulation, and the downregulation of this action is consistent with the idea that limiting ribosome production is beneficial during the stress response. Although many processes have been

implicated in the general response, the additional principal components offer an explanation for what may be unique to each stress state.

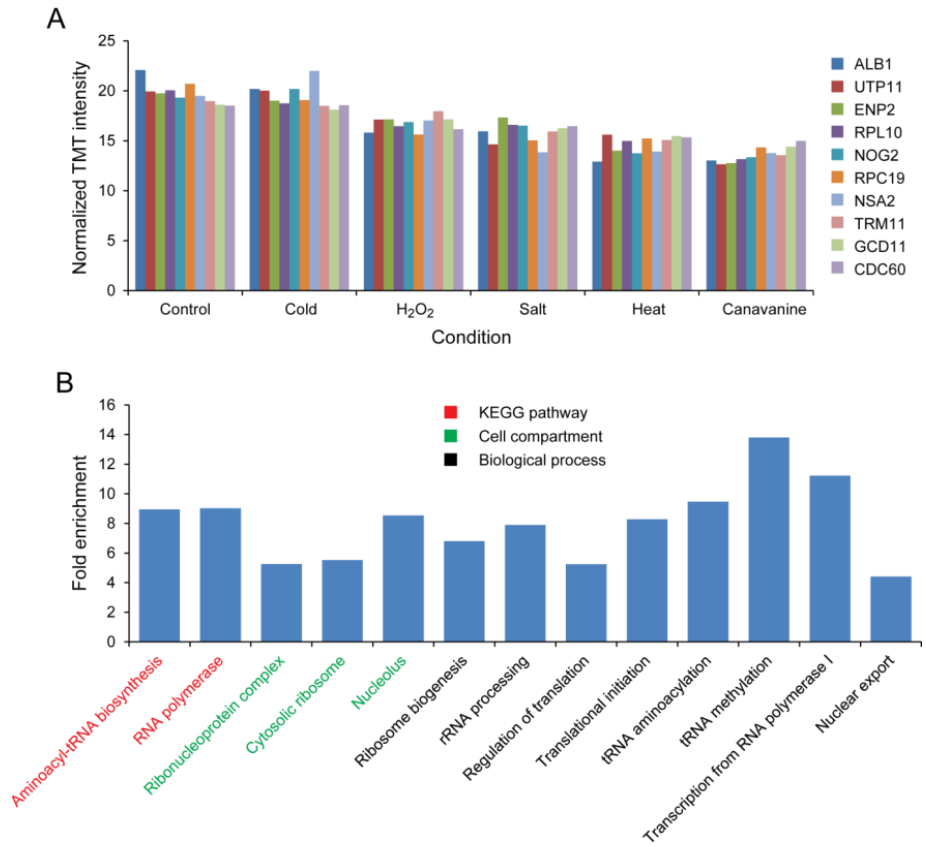


Figure 4.61. Gene ontology analysis of the general, downregulated stress response. (A) Highlighted hits from the top 100 principal component one loading values (positive direction), and (B) associated gene ontology. As with the upregulated general response, the downregulated proteins are generally those observed in the heat stress response. It is clear that the downregulation of ribosomal machinery may be important to all stress states.

Component two may explain responses which are more specific to heat stress. As before the top 100 principal component loadings were analyzed for gene ontology (Figure 4.62). Interestingly ER luminal proteins (~25 fold), and ER-associated protein catabolism (~8 fold) were enriched. These may be indicative of another level of the protein misfolding response, namely the ERAD pathway. As previously found, NAD metabolic categories (~20 fold enrichment) and arginine metabolism (~20 fold enrichment were specifically found in the heat stress component, further solidifying their role. Additional protein folding components were observed as well (~ 4 fold enrichment).

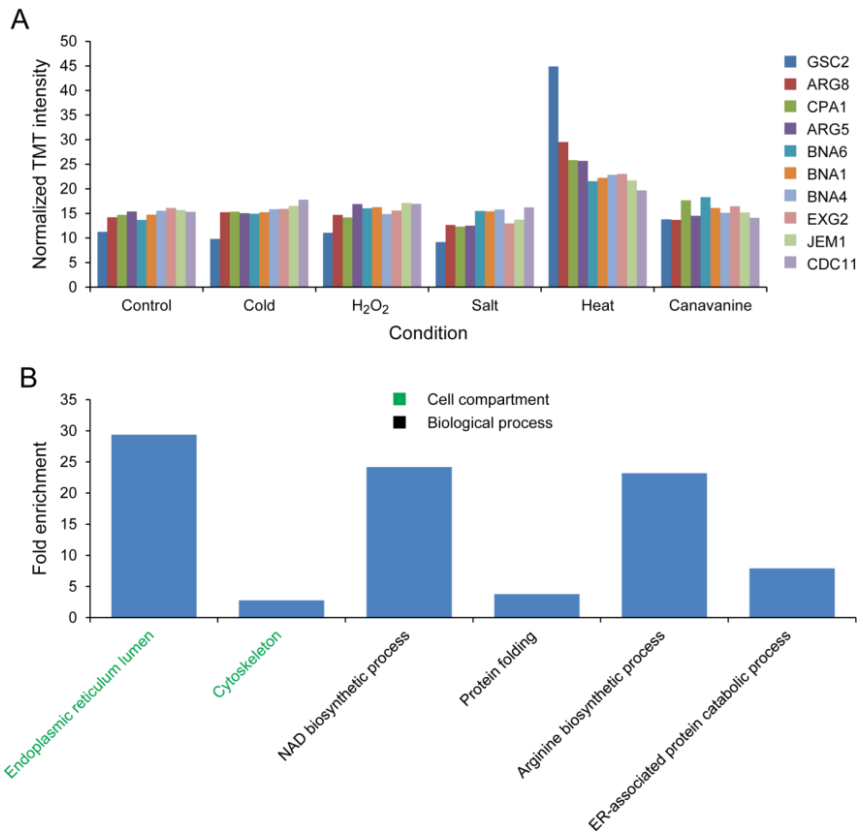


Figure 4.62. Gene ontology analysis of the heat stress specific component. (A) Highlighted hits from the top 100 principal component two loading values (negative direction), and (B) associated gene ontology. Additional categories above the general stress response are added in this component including those involved in NAD metabolism, arginine biosynthesis, and the ERAD pathway.

As highlighted, component three greatly separated oxidative stress from the other stress (negative direction, figure). Gene ontology results for the top 100 proteins contained in this component are summarized in Figure 4.63. Not surprisingly, many of the proteins found in component three were mitochondrial (lumen/matrix and intermembrane space were enriched ~4 and ~6 fold, respectively), consistent with its role in the response to ROS. Processes involved specifically with the reaction to ROS (~5 fold) and superoxide (~18 fold) were enriched. Consistent with these observations, the Kegg pathway for glutathione metabolism, an important mediator of redox homeostasis, is enriched (~10 fold). Proteins identified in this component were also enriched with the mitochondrial ribosome and mitochondrial translation categories. As the main function of mitochondrial transition is the production

of proteins involved in oxidative phosphorylation, it may indicate that oxidative stress disrupts normal oxidative phosphorylation, and yeast compensate through the upregulation of complex components. Alternatively, it may reflect an overall increase in aerobic metabolic activity. The upregulation of mitochondrial specific translation contrasts the general downregulation of cytosolic ribosomal components. In addition to metabolic adaptation, proteasomal components were enriched (~10 fold). An increase in protein catabolism may be a response to protein misfolding, which was also proposed to occur upon H_2O_2 treatment.

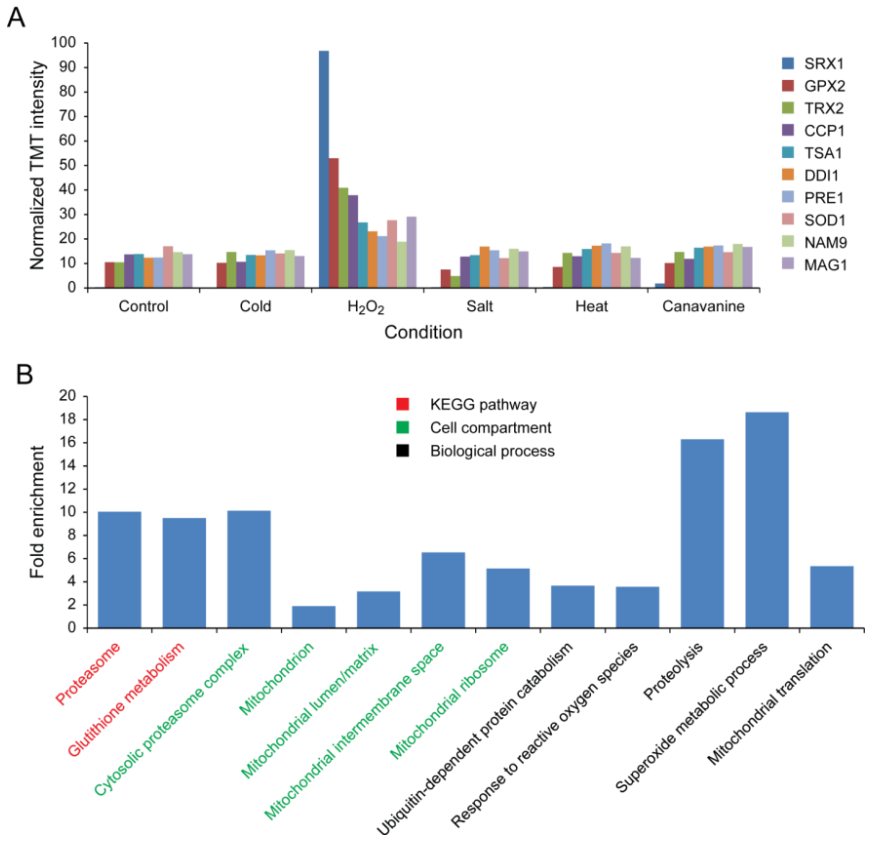


Figure 4.63. Gene ontology analysis of the oxidative stress specific component. (A) Highlighted hits from the top 100 principal component three loading values (negative direction), and (B) associated gene ontology. Additional categories above the general stress response are added in this component including those involved redox metabolism (glutathione metabolism, response to ROS and superoxide), protein catabolic processes (proteasome, proteolysis, ubiquitin dependent protein catabolism) and mitochondrial translation (including mitochondrial ribosome components). Many of the proteins found in this component were mitochondrial.

A direct analysis of either component three (positive direction) or component four did not reveal any significant gene ontology categories. As both components were implicated in the salt stress response, they loading values for components three and four (both positive directions) were summed and sorted by that summed value. The top 100 proteins from this list were used for the analysis of salt stress specific responses (Figure 4.64). Not surprisingly proteins previously annotated in the osmotic response were contained within this group (~ 4 fold enriched). Glycerol metabolism was highly enriched in this group of proteins (~20 fold), a known response to osmotic stress. This group of proteins was also enriched with both the cell cortex and cytoskeletal components (~ 5 fold), suggesting responses in structural organization, particularly at the plasma membrane may be involved with response to osmotic stress. Consistent with this observation, categories of membrane origination, endocytosis and vesicular trafficking were enriched (~ 3-5 fold), suggesting the regulation of membrane dynamics may be particularly important to osmotic stress adaptation. Additional proteins involved in carbohydrate catabolism (~5 fold enriched) were observed.

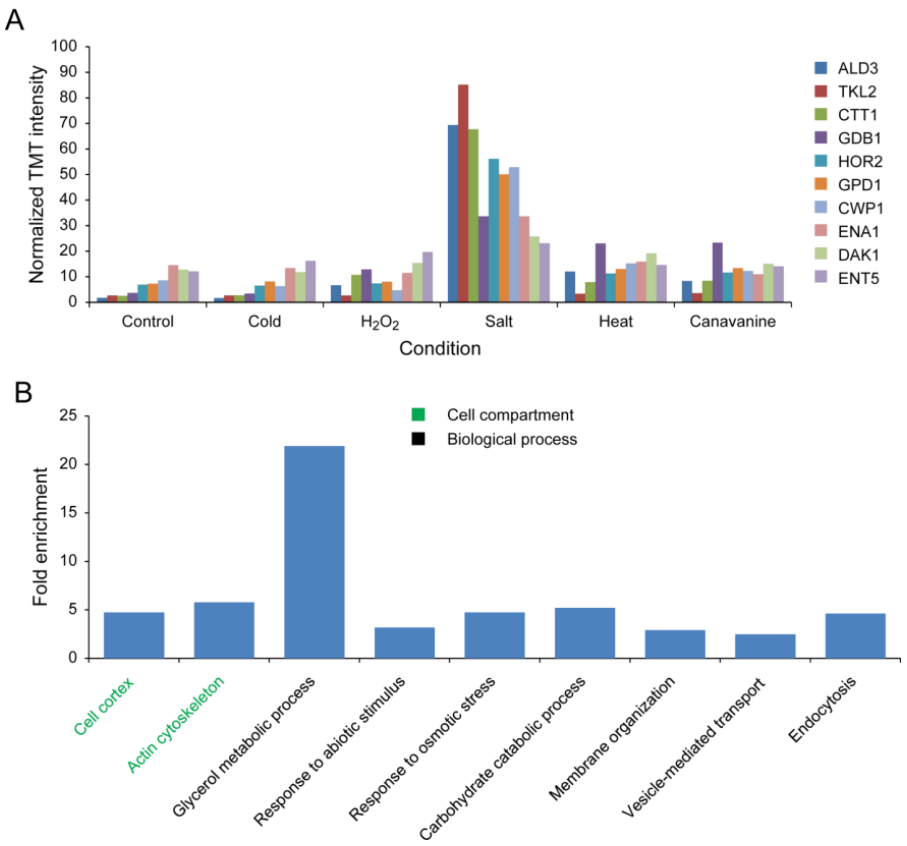


Figure 4.64. Gene ontology analysis of the salt stress specific components. Principal component 3 (positive direction) and principal component four (positive direction) alone were not sufficient to separate out salt stress from the other stresses (by significant gene ontology category identification), therefore the sum of PC3 and PC4 was used in the analysis. (A) Highlighted hits from the top 100 principal component three and principal component four loading values (summed, positive direction), and (B) associated gene ontology. In addition to the general stress response, categories specific to salt stress were identified. The majority of new categories seem to be involved with membrane regulation, including the cell cortex and cytoskeleton. Membrane organization, endocytosis and vesicle-mediated transport further implicate the plasma membrane in the salt stress response. Glycerol metabolism, a known mediator of osmotic stress response was uncovered.

The Use of Non-Negative Matrix Factorization (NMF) for the Analysis of Five Yeast Stress States

It is clear from the principal component analysis that both general and unique stress adaption exists. The general response defined by component one, represents the largest responses observed. In comparison, the more specific stress adaptations tend to involve fewer proteins (less explained variance), but are more tailored to the exact needs of yeast under that condition. Though PCA was useful in defining the underlying biological processes in the stress response, as discussed, drawing

specific cutoffs for the proteins involved in such process can be arbitrary. To specifically define these proteins, NFM was applied to the analysis of these stress states

The first step in the NMF analysis of the discussed stress states was to determine appropriate number of clusters (k). As five stresses were compared to a control, where four of the five stresses demonstrated considerable protein responses, it was reasonable to assume that ideal number of cluster would be five. As before, clusters obtain using k=2 through k=6 were analyzed for their consistency (Figure 4.65). The most consistent clustering of the stress conditions into consensus groups was achieved with five clusters, followed closely by six. Using five clusters, H_2O_2, salt, heat and canavanine replicates clustered with one another, whereas the remaining control and cold samples cluster together. In support of the use five clusters, the samples always showed the same cluster pattern, independent of the random starting point of an NMF iteration. Using six clusters, the longer cold stress sample showed a slight preference for grouping with the shorter cold point, suggesting some changes did occur in these samples; as stated it may be of use to conduct a long (~24 hr) cold stress time course to determine the correct temporal pattern of regulation at low temperatures. In support of using five clusters for the analysis, the cophenetic score was maximized at K = 5 (Figure 4.66). The inflection point in the residuals (and rss) was also observed at K =5, supporting that the use of five clusters minimizes over fitting.

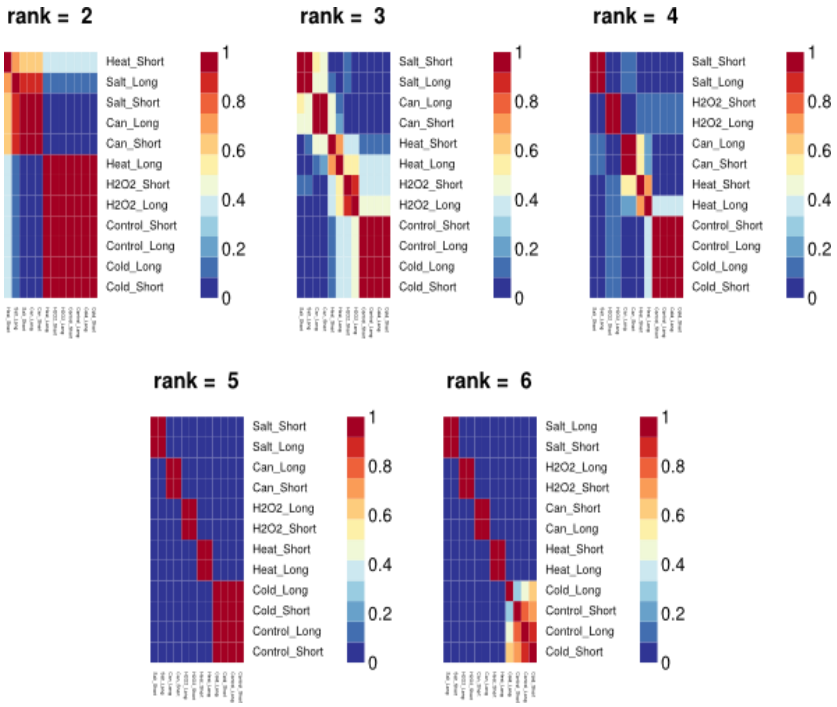


Figure 4.65. Reordered consensus maps of the five stress data, using different number of clusters (K=2 to K=6). The average consensus map over all NMF iterations is presented for each value, k. The color bar represents the frequency in which two samples cluster together throughout the iterations. A rank of K = 5 showed the most consistent clustering, suggesting five groups of proteins exist in the data set, and is consistent with the best consensus map found among all NMF iterations.

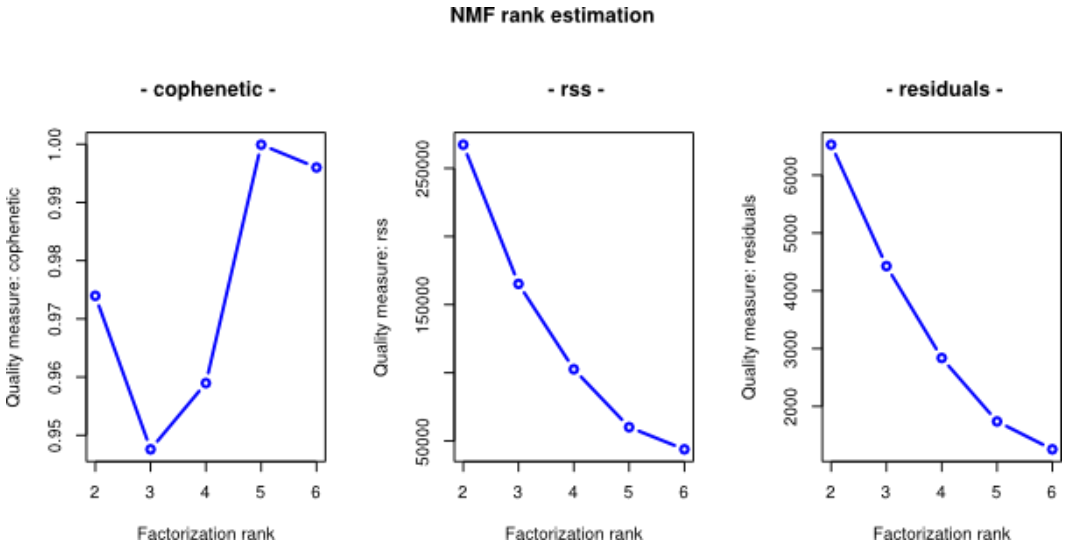


Figure 4.66. Rank estimation for five stress data set, based on cophenetic correlation and the residual sum of squares/residual values. As with the time course NMF analysis, the maximal cophenetic value is correlated with stable clustering, as is the point of inflection with the residual sum of squares and residual values. Both metrics support the use five clusters. In addition, since only four of the five stresses caused significant protein regulation (cold stress was very similar to the control), five clusters is consistent with the expected biology (four stress groups and one control group). As before, other common metrics, such as dispersion, also support the use of five clusters.

The best connectivity matrix (containing the least error among all NMF iterations),

corresponding to the basis and coefficient matrices used for further analysis is displayed in Figure 4.67.

The basis number (also corresponding to the protein group number) is displayed, as is the consensus number. As discussed this matrix contains five consensus groups, one for each of the four stressed demonstration significant changes, and one for the cold and control samples.

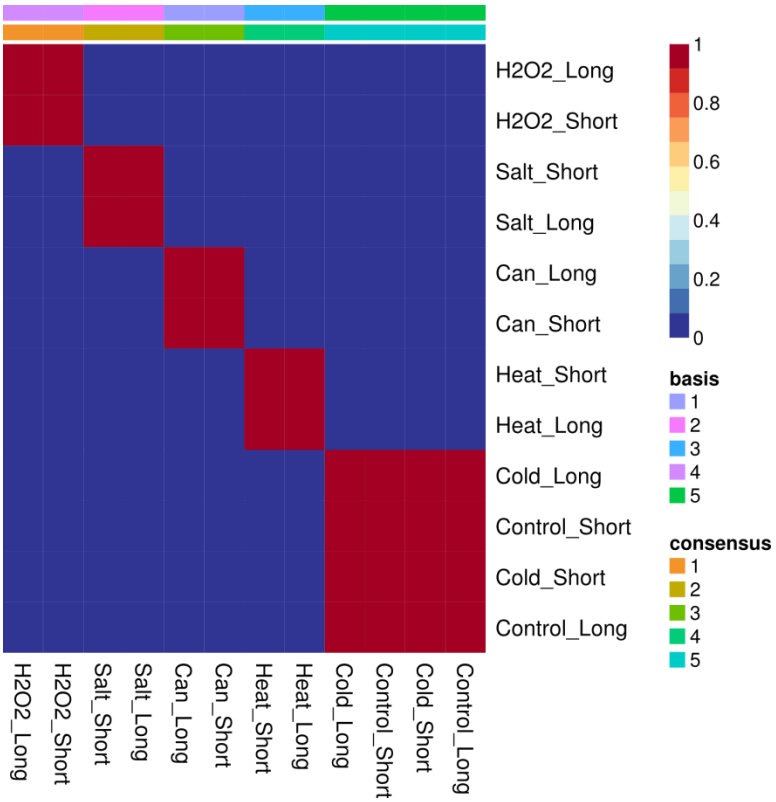


Figure 4.67. Best clustering consensus among NMF iterations. The heat map of the best connectivity matrix, indicating basis number and consensus groups, is displayed. The color values are based on which consensus the samples fall within. The map indicates that five consistent groups exist within the data, each of which represents a stress state, with the exception of cold stress. The basis and coefficient matrices associated with this consensus map were used in further operations, such as feature extraction for identifying proteins which represent each group

As described above for the heat stress time course, the coefficient matrix contains information about the relationship between biological samples and the clusters, including the stability with which a given sample contributes to a cluster (Figure 4.68). Generally only one stress contributed to each basis, with the exception of basis five; the cold stress, control, and to some degree the heat stress, contributed to that basis. Basis four was the most specific, and was contributed to by H_2O_2 stress fairly exclusively. Among the stresses that changed, basis three was the least specific. Though it was primarily influenced by heat stress, and thus represents proteins which are more uniquely regulated in that stress, canavanine and salt stresses also contributed to this basis. Canavanine stress also contributed to basis two to some degree (salt stress basis); this observation that canavanine contributed to many groups is

consistent with the PCA analysis. Though there was some overlap in basis contribution between the stresses, each one generally was specific to a stress state, and representative proteins were extracted as with the heat stress time course.

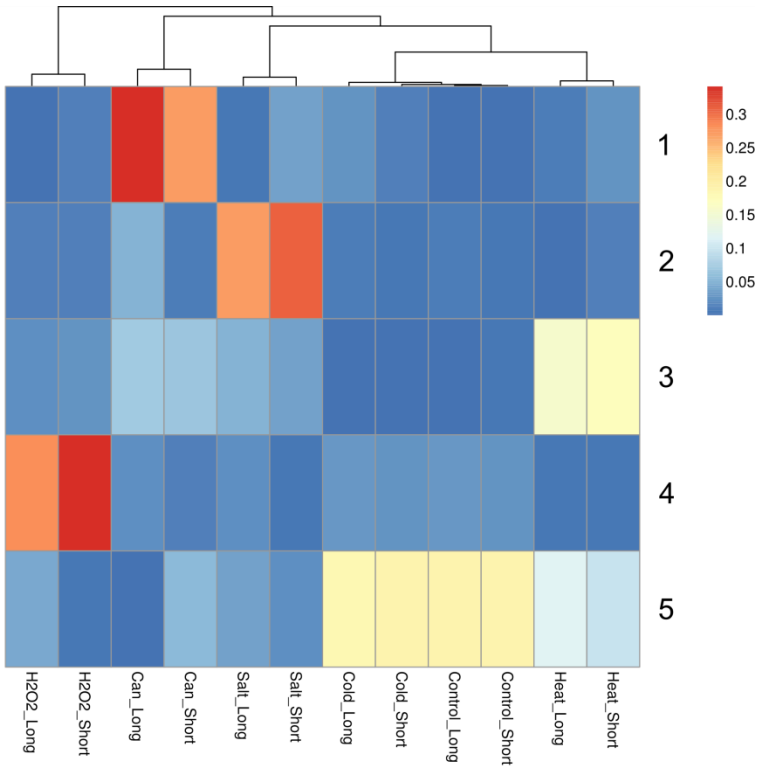


Figure 4.68. Five stresses NMF coefficient matrix. A hierarchical cluster of coefficient values is displayed on top of the diagram, as are the basis numbers on the right. With the exception of the cold stress, the stresses each were grouped into their own basis by replicate (1, canavanine; 2, salt; 3, heat; 4, H_2O_2). The control stress replicates and the control replicates were grouped into basis five. The basis clustering and the hierarchical clusters (top of heat map) agreed. The canavanine stress and salt stress contributed to the heat stress basis, suggesting there are some shared stress responsive proteins within these stresses. Heat stress also contributed to the control basis, suggesting there are proteins which are regulated in other stresses which are not regulated in heat stress.

Expression profiles of the extracted features of each basis follow the coefficient heat map (Figure 4.69). Each group of proteins is either primarily regulated, or in some cases, uniquely regulated in a given stress. Group one represents proteins upregulation in canavanine stress; group two contains those proteins upregulated in salt stress; group three embodies those upregulated in heat stress; group four represents proteins upregulation upon H_2O_2 stress; finally, group five contains proteins which are

downregulated amongst all the stresses. No group of stress-specific downregulated proteins was found, suggesting it is a general process. This general pattern of downregulated follows the trend observed in the PCA analysis; namely that the general response is most significant in the canavanine stress, followed by heat stress, and then equally salt and H_2O_2 stresses. Interestingly, as with the group of downregulated proteins in basis one from the heat stress time course, NMF seemed to treat downregulated proteins as proteins upregulated in the control (maximal coefficient values in the control samples). This result demonstrates the blind nature of NMF, that it does not assume any temporal or biological relationships among the samples.

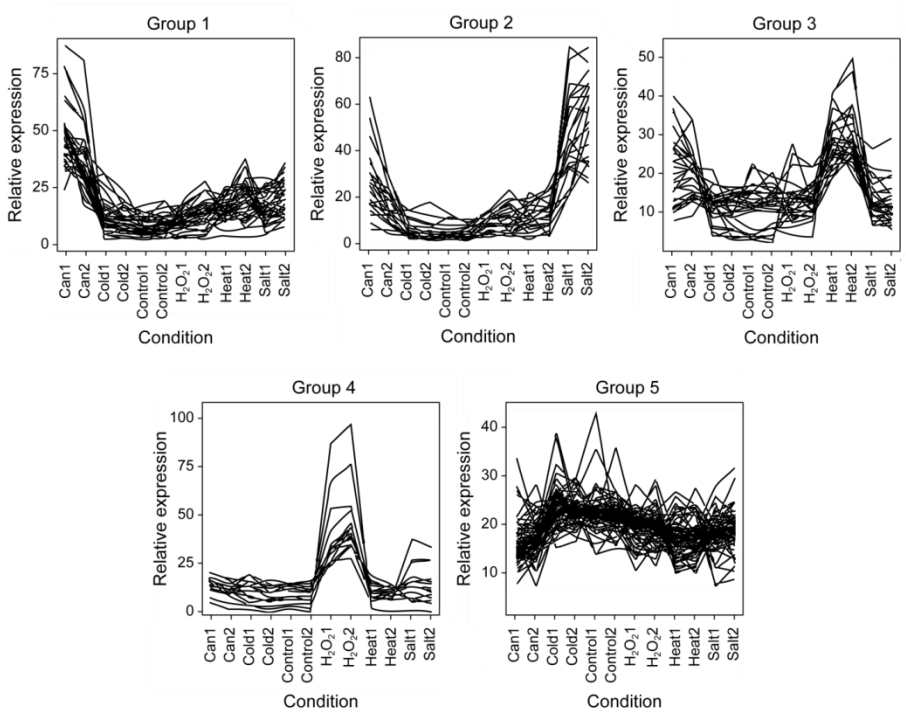


Figure 4.69. Expression profiles for extracted protein groups from basis 1-5. Group one is primarily canavanine stress ("Can"), group two is primarily salt stress, group three is primarily heat stress, group four is nearly exclusively H_2O_2 stress, and group five are commonly downregulated proteins. Each group represents proteins which are for the most part primarily regulated in one stress; a protein may be upregulated in multiple stresses, but a visible stress-specific magnitude of expression is visible in all groups. That being said, there are also commonalities between stresses as well, such as the pattern displayed in group 3. Though the protein expression pattern of group three is most dominant in heat stress, due to the related biology of canavanine treatment and heat stress, a spike in the expression profile is also observed in canavanine stress. The behavior displayed here exemplifies previously discussed behavior, namely that there is a large general stress response, and more subtle stress-specific responses, in the context of the discussed stresses.

The proteins which represent each stress state are listed below (Table 4.5). Though some of the highlighted general stress responder discussed in the PCA were found in the canavanine group again (e.g. HXK1, SSA4), NMF generated a more unique, albeit smaller, list of proteins representative of the canavanine response. Many of the highlighted stress specific proteins in the salt, heat, and H_2O_2 stresses found by PCA were again found by NMF. Due to the more specific lists of proteins, gene ontology p-values were lower with the NMF analysis in some cases, compared to PCA. For example, the p-value for arginine biosynthesis (heat stress) by PCA analysis was 0.04 (just below the cutoff), while by NMF analysis it was 0.0005 (well below the cutoff). Thus NMF may be able to, in some cases, extract significant gene ontology categories, which may have not otherwise been found through PCA. In addition, the concise group of proteins generated by NMF allows a more specific avenue for future biological experimentation. NMF and PCA are both complementary and confirmational, with respect to one another, and should both be considered for the analysis of proteomics data.

Table 4.5. Extracted groups of proteins from each basis in the five stress experiment. Features were extracted using a pre-defined scoring algorithm, as before. Proteins in each group are those which help define a given stress by NMF. Group one represents canavanine stress (N = 29), group two represent salt stress (N = 20), group three represent heat stress (N = 42), group four represent H_2O_2 stress (N = 15), and group five represents common downregulation (N = 140). No group of stress specific downregulated proteins were found, suggesting downregulation is very general.

Group 1

[1] BTN2 SSA3 HSP26 SSA4 PRY1
[6] FMP16 HOR7 HSP78 HSP12 PCL1
[11] HXK1 RPN4 YBR085C-A SML1 HSP42
[16] NCE103 GCY1 SED1 DCS2 PHM8
[21] SOL4 NRG1 STF2 FES1 YGR250C
[26] PIC2 HSP150 AIM17 YGR127W

Group 2

[1] YHR033W TKL2 NQM1 FMP45 CTT1
[6] ALD3 HBT1 GND2 HOR2 PHM7
[11] RTN2 BDH2 CWP1 GPD1 GLO1
[16] MSC1 RTC3 PAI3 NCE102 ARO9

Group 3

[1] GSC2 ADH2 GPM2 AHA1 HSP82
[6] YDR222W ARG8 YNR034W-A HXT7 ARG5
[11] CRH1 HCH1 YER067W HSP104 TNA1
[16] TDH1 GPH1 CPA1 APJ1 ARG1
[21] HSP60 GDB1 ARO10 PGM2 CPR6
[26] HSP10 SSA1 TFS1 ALD4 PTR2
[31] BNA4 PRY2 YOR142W-B YPL067C TSL1
[36] GAD1 PBI2 HSC82 UTH1 GRX1
[41] STI1 YMR196W

Group 4

[1] SRX1 ECM4 HBN1 YMR090W GPX2
[6] TRX2 CYC1 YML131W YLR108C YLR460C
[11] ISU2 PRX1 TRR1 GRE2 CCP1

Group 5

[1] STE3 BUD20 ZRT1 RLP24 FPR1
[6] HXT2 GIS2 RPL34B SAM3 HXT3
[11] IZH2 REX4 TMA16 RRP8 DBP2
[16] OLE1 FAR1 ASH1 PDR12 TMA7
[21] INO1 SAM2 YBR141C SAM1 SSB2
[26] MFα1 GRX8 YBL028C IRC5 YHB1
[31] RMT2 QDR2 NSA2 REX3 MDH2
[36] ALT2 RNH70 ACB1 SST2 CBC2
[41] KEX2 IMD2 GFD2 SBH2 ALB1
[46] MET6 SCP1 HXT1 EXG1 DOT6
[51] HHT1 NOP4 FYV4 NOP15 GAS3
[56] HIP1 PXR1 ATF2 RRP15 MUP1
[61] CIC1 RPL33B RPF1 MAK16 YOL019W
[66] NOG2 RPL43B RPS10B BUD22 NSR1
[71] RPS8A RPS16B TRX1 ZEO1 RNQ1
[76] NHP2 RPS28B NOP16 HMT1 PUS7
[81] RPS29B PHO3 RRS1 CDC36 FEN1
[86] RPA14 TRM2 YDR514C CPT1 SPO12
[91] URA1 RNH202 RPL15A RPS26B MRD1
[96] GNP1 DBP3 GDT1 RPL19B SRO77
[101] CAN1 RPL32 LRP1 POP6 HHO1
[106] MET10 IMP3 FYV7 RPL16B YAL044W-A
[111] SAS10 YMD8 NHP6A VTS1 RPL20A
[116] UTR2 UTP11 RPS4B NAF1 NIS1
[121] CPR8 SDS23 NOP6 SRP21 TIF35
[126] RPL16A EAR1 YDR262W NUG1 ENP2
[131] TOS1 RPS18A NOP2 UTP14 MHT1
[136] HTB2 YGL101W SKG6 RRP14 NOG1

The Future of Quantitative Multiplexing for Proteomic Analysis

Ultimately, it was demonstrated in this chapter that proteomics has progressed to the point that it is competitive with genomics, in terms of depth (within the yeast proteome), and analytical ability (multiplexing capability and accuracy). Indeed many of the resources which were previously available only to the genomics community are now applicable in the analysis of proteomics data sets. Often many genomics data sets are combined to look for common and unique features amongst a variety of samples, such as the comparison of multiple cancer cell lines or even primary tumors. To asses such

feasibility for proteome level comparison, all data set were combined, based on their normalized TMT (relative abundance of TMT ions in each channel) within one experiment. These data were hierarchically clustered, and the results are presented below (Figure 4.70). Interestingly, the data grouped by the stress state and not by the experiment from which they were derived. There is an obvious control group, which contains the controls from all experiments and the cold stresses (highlighted in blue, Figure 4.70). There is also a heat stress group (highlighted in red, Figure 4.70); within this group, the time points of heat stress analysis tended to group together. For example the 60 min time point from the five stress experiment and the 60 minute time point from the heat stress time course grouped together. The triplicate heat stress samples did not cluster with the other 60 min points, though they were still within the general heat stress cluster. This last point highlights the need for a proper normalization between experiments, to account for difference in experimental design and technical variability. A means of normalization may be the inclusion of an identical common sample within all experiments.

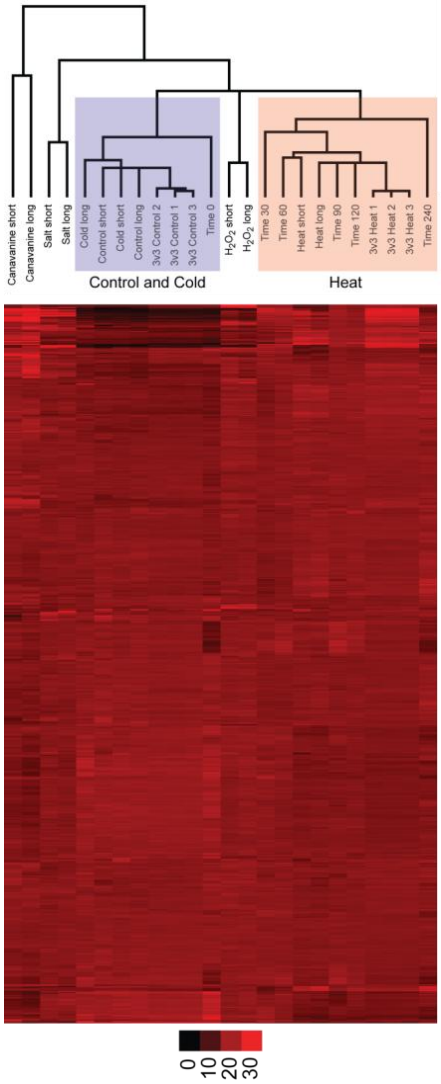


Figure 4.70. Stress data cluster by their treatment condition when all data sets are combined. Only proteins which were quantified in all experiments were clustered. The color bar represents the % percent of TMT signal in a given channel within one set of six conditions. All stress conditions group with one another, and not by experiment. A group of control samples (including the cold stresses) and heat stresses are highlighted in blue and red, respectively. The time points within the heat stress generally cluster together, such as the 60 min point from the 5 stresses experiment and the 60 min point from the heat stress time course. The triplicate heat stress triplicate samples (as with the control triplicates) did not cluster with the other 60 min time points, though were still contained within the general heat stress group. Though these facts highlight the use of TMT for comparing multiple experiments simultaneously, it is evident that proper normalization, such as though the inclusion of a common sample in all experiment, is required for the most accurate analysis.

Despite some of the issues regarding normalization between experiments, within this large clustering array, specific groups of proteins were still identified. Groups of proteins which were primarily upregulated in oxidative stress (Figure 4.71, A), generally upregulated across all stresses (Figure 4.71, B), primarily upregulated in canavanine and heat stresses (Figure 4.71, C), primarily upregulated in salt stress (Figure 4.71, D), primarily upregulated in heat stress (Figure 4.71, E), and generally downregulated across all stresses (Figure 4.71, F) were identified. With proper normalization, such groups may be further defined, and new groups may even be extracted. The ability for such normalization has recently

been achieved. TMT reagents which allow up to 10-plex experimentation are available, which would allow the biological triplicate analysis of three conditions at once, leaving one channel available for use as this common sample between experiments. Combined with metabolic incorporation of stable isotopes (light, medium and heavy SILAC) [43], one could perform nine biological triplicate analyses in a single LC-MS experiment. Quantitative multiplexing has redefined what is possible in proteomics, and has left the future open to unprecedented biological analysis.

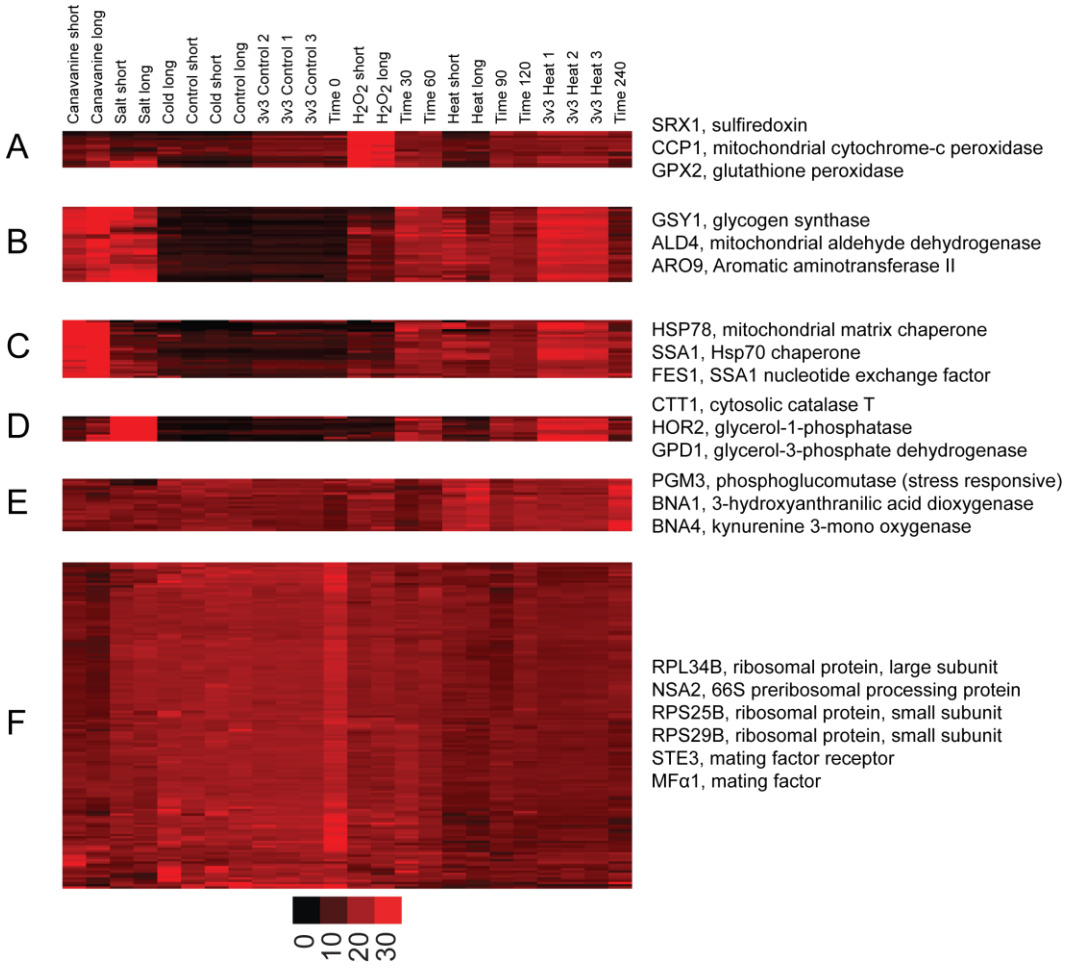


Figure 4.71. Groups of stress specific and common stress response proteins are still identifiable when clustering all data sets at once. Each group represents the following conditions primarily: (A) H_2O_2 stress, (B) general upregulated stress response, (C) canavanine and heat stresses (greater expression in canavanine stress), (D) salt stress, (E) heat stress, and (F) general downregulated stress response. Highlight proteins and their annotations are indicated.

Conclusions

Here I demonstrated technological improvements which permit the accurate quantification of thousands of proteins, from multiple conditions simultaneously, and applied the improved methods to an analysis of the yeast stress response. HPRP pre-fractionation was demonstrated to be a robust method, which produced consistently rich LC-MS fraction through the combination of early, mid, and late HPRP fractions. This method was superior to SCX in terms of protein coverage (unique peptide and protein identifications), and required less total fractions and analysis time to achieve these gains. The multinotch method was demonstrated to increase TMT reporter ion signal, leading to more accurate quantification, while still avoiding interference; the result of which is a greater number of relevantly quantified proteins. The combination of these methods yields a robust strategy for proteome wide quantitative multiplexing.

Three common multiplexing experiments were highlighted: a biological triplicate analysis of heat stress, a heat stress time course, and a multi stress state comparison. In the yeast heat stress response, hundreds of proteins were found to be significantly regulated by T-test, using biological replicates. The time course measurements revealed that heat stress responsive proteins could be grouped into consistently up and downregulated temporal patterns, or a transiently regulated pattern; the transiently regulated proteins could further be divided into additional categories. Multi stress state comparisons revealed both a common stress response, and those more unique to the stresses tested (H_2O_2, salt, heat and canavanine). In all experiments the use of dimensionality reduction was highlight, through methods such as PCA and NMF. The biological response observed during stress took the general form of a starvation response, in which many catabolic pathways were upregulated. Interestingly some anabolic pathways, such as arginine synthesis in heat stress, were found to be upregulated. In many cases some paradoxical regulation was observed, such as the simultaneous upregulation of anabolic and catabolic

branches of the glycogen pathway. Clearly the stress response is a complex system, and these data offer a context for future endeavors. This chapter provides a framework for the acquisition of accurate multiplexed data sets, and demonstrates the bioinformatic tools which may be applied to the data, in order to extract biologically relevant information. This analysis obviated the increasing need for replicates in large scale biology, an important consideration for future experimental design.

References

1. Sanger, F.; Coulson, A. R., A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **1975,** 94, (3), 441-8.
2. Sanger, F.; Air, G. M.; Barrell, B. G.; Brown, N. L.; Coulson, A. R.; Fiddes, C. A.; Hutchison, C. A.; Slocombe, P. M.; Smith, M., Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **1977,** 265, (5596), 687-95.
3. Sanger, F.; Nicklen, S.; Coulson, A. R., DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **1977,** 74, (12), 5463-7.
4. Church, G. M.; Kieffer-Higgins, S., Multiplex DNA sequencing. *Science* **1988,** 240, (4849), 185-8.
5. Khoury, M. J.; McBride, C. M.; Schully, S. D.; Ioannidis, J. P.; Feero, W. G.; Janssens, A. C.; Gwinn, M.; Simons-Morton, D. G.; Bernhardt, J. M.; Cargill, M.; Chanock, S. J.; Church, G. M.; Coates, R. J.; Collins, F. S.; Croyle, R. T.; Davis, B. R.; Downing, G. J.; Duross, A.; Friedman, S.; Gail, M. H.; Ginsburg, G. S.; Green, R. C.; Greene, M. H.; Greenland, P.; Gulcher, J. R.; Hsu, A.; Hudson, K. L.; Kardia, S. L.; Kimmel, P. L.; Lauer, M. S.; Miller, A. M.; Offit, K.; Ransohoff, D. F.; Roberts, J. S.; Rasooly, R. S.; Stefansson, K.; Terry, S. F.; Teutsch, S. M.; Trepanier, A.; Wanke, K. L.; Witte, J. S.; Xu, J., The Scientific Foundation for personal genomics: recommendations from a National Institutes of Health-Centers for Disease Control and Prevention multidisciplinary workshop. *Genet Med* **2009,** 11, (8), 559-67.
6. Hardenbol, P.; Yu, F.; Belmont, J.; Mackenzie, J.; Bruckner, C.; Brundage, T.; Boudreau, A.; Chow, S.; Eberle, J.; Erbilgin, A.; Falkowski, M.; Fitzgerald, R.; Ghose, S.; Iartchouk, O.; Jain, M.; Karlin-Neumann, G.; Lu, X.; Miao, X.; Moore, B.; Moorhead, M.; Namsaraev, E.; Pasternak, S.; Prakash, E.; Tran, K.; Wang, Z.; Jones, H. B.; Davis, R. W.; Willis, T. D.; Gibbs, R. A., Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res* **2005,** 15, (2), 269-75.
7. Syka, J. E.; Marto, J. A.; Bai, D. L.; Horning, S.; Senko, M. W.; Schwartz, J. C.; Ueberheide, B.; Garcia, B.; Busby, S.; Muratore, T.; Shabanowitz, J.; Hunt, D. F., Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J Proteome Res* **2004,** 3, (3), 621-6.
8. Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M., Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **2002,** 1, (5), 376-86.
9. Wenger, C. D.; Lee, M. V.; Hebert, A. S.; McAlister, G. C.; Phanstiel, D. H.; Westphall, M. S.; Coon, J. J., Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nat Methods* **2011,** 8, (11), 933-5.

10. Ting, L.; Rad, R.; Gygi, S. P.; Haas, W., MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods* **2011,** 8, (11), 937-40.
11. Kim, H.; Park, H., Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* **2007,** 23, (12), 1495-502.
12. Brunet, J. P.; Tamayo, P.; Golub, T. R.; Mesirov, J. P., Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* **2004,** 101, (12), 4164-9.
13. McAlister, G. C.; Jedrychowski, M.; Yu, Y.; Ting, L.; Huttlin, E. L.; Rad, R.; Haas, W.; Gygi, S. P. In *Isolating multiple MS2 fragments using waveforms with multiple frequency notches improves MS3 sensitivity ~8 fold over standard MS3-based TMT methods.*, 60th Annual Conference of the American Society of Mass Spectrometry, Vancouver, Canada, May 20−24, 2012; 2012.
14. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007,** 4, (3), 207-14.
15. Huttlin, E. L.; Jedrychowski, M. P.; Elias, J. E.; Goswami, T.; Rad, R.; Beausoleil, S. A.; Villen, J.; Haas, W.; Sowa, M. E.; Gygi, S. P., A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **2010,** 143, (7), 1174-89.
16. Benjamini, Y.; Drai, D.; Elmer, G.; Kafkafi, N.; Golani, I., Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* **2001,** 125, (1-2), 279-84.
17. Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D., Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **1998,** 95, (25), 14863-8.
18. Gaujoux, R.; Seoighe, C., A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **2010,** 11, 367.
19. Smyth, G. K., Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **2004,** 3, Article3.
20. Dennis, G., Jr.; Sherman, B. T.; Hosack, D. A.; Yang, J.; Gao, W.; Lane, H. C.; Lempicki, R. A., DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **2003,** 4, (5), P3.
21. Mostafavi, S.; Ray, D.; Warde-Farley, D.; Grouios, C.; Morris, Q., GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol* **2008,** 9 Suppl 1, S4.
22. Montojo, J.; Zuberi, K.; Rodriguez, H.; Kazi, F.; Wright, G.; Donaldson, S. L.; Morris, Q.; Bader, G. D., GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics* **2010,** 26, (22), 2927-8.
23. Gasch, A. P.; Spellman, P. T.; Kao, C. M.; Carmel-Harel, O.; Eisen, M. B.; Storz, G.; Botstein, D.; Brown, P. O., Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **2000,** 11, (12), 4241-57.
24. Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R., Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **1999,** 19, (3), 1720-30.
25. Lee, M. V.; Topper, S. E.; Hubler, S. L.; Hose, J.; Wenger, C. D.; Coon, J. J.; Gasch, A. P., A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Mol Syst Biol* **2011,** 7, 514.
26. Wu, R.; Dephoure, N.; Haas, W.; Huttlin, E. L.; Zhai, B.; Sowa, M. E.; Gygi, S. P., Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes. *Mol Cell Proteomics* **2011,** 10, (8), M111 009654.
27. Francois, J.; Parrou, J. L., Reserve carbohydrates metabolism in the yeast Saccharomyces cerevisiae. *FEMS Microbiol Rev* **2001,** 25, (1), 125-45.
28. Wiemken, A., Trehalose in yeast, stress protectant rather than reserve carbohydrate. *Antonie Van Leeuwenhoek* **1990,** 58, (3), 209-17.

29. Kabani, M.; Beckerich, J. M.; Brodsky, J. L., Nucleotide exchange factor for the yeast Hsp70 molecular chaperone Ssa1p. *Mol Cell Biol* **2002,** 22, (13), 4677-89.
30. Lopez, N.; Halladay, J.; Walter, W.; Craig, E. A., SSB, encoding a ribosome-associated chaperone, is coordinately regulated with ribosomal protein genes. *J Bacteriol* **1999,** 181, (10), 3136-43.
31. Pfund, C.; Lopez-Hoyo, N.; Ziegelhoffer, T.; Schilke, B. A.; Lopez-Buesa, P.; Walter, W. A.; Wiedmann, M.; Craig, E. A., The molecular chaperone Ssb from Saccharomyces cerevisiae is a component of the ribosome-nascent chain complex. *EMBO J* **1998,** 17, (14), 3981-9.
32. Moehle, C. M.; Hinnebusch, A. G., Association of RAP1 binding sites with stringent control of ribosomal protein gene transcription in Saccharomyces cerevisiae. *Mol Cell Biol* **1991,** 11, (5), 2723-35.
33. Lopez, M. C.; Baker, H. V., Understanding the growth phenotype of the yeast gcr1 mutant in terms of global genomic expression patterns. *J Bacteriol* **2000,** 182, (17), 4970-8.
34. Chern, M. K.; Chang, K. N.; Liu, L. F.; Tam, T. C.; Liu, Y. C.; Liang, Y. L.; Tam, M. F., Yeast ribosomal protein L12 is a substrate of protein-arginine methyltransferase 2. *J Biol Chem* **2002,** 277, (18), 15345-53.
35. Tuorto, F.; Liebers, R.; Musch, T.; Schaefer, M.; Hofmann, S.; Kellner, S.; Frye, M.; Helm, M.; Stoecklin, G.; Lyko, F., RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. *Nat Struct Mol Biol* **2012,** 19, (9), 900-5.
36. Welch, A. Z.; Gibney, P. A.; Botstein, D.; Koshland, D. E., TOR and RAS pathways regulate desiccation tolerance in Saccharomyces cerevisiae. *Mol Biol Cell* **2013,** 24, (2), 115-28.
37. Schule, T.; Rose, M.; Entian, K. D.; Thumm, M.; Wolf, D. H., Ubc8p functions in catabolite degradation of fructose-1, 6-bisphosphatase in yeast. *EMBO J* **2000,** 19, (10), 2161-7.
38. Regelmann, J.; Schule, T.; Josupeit, F. S.; Horak, J.; Rose, M.; Entian, K. D.; Thumm, M.; Wolf, D. H., Catabolite degradation of fructose-1,6-bisphosphatase in the yeast Saccharomyces cerevisiae: a genome-wide screen identifies eight novel GID genes and indicates the existence of two degradation pathways. *Mol Biol Cell* **2003,** 14, (4), 1652-63.
39. Xie, Z.; Nair, U.; Klionsky, D. J., Atg8 controls phagophore expansion during autophagosome formation. *Mol Biol Cell* **2008,** 19, (8), 3290-8.
40. Onodera, J.; Ohsumi, Y., Autophagy is required for maintenance of amino acid levels and protein synthesis under nitrogen starvation. *J Biol Chem* **2005,** 280, (36), 31582-6.
41. He, C.; Klionsky, D. J., Regulation mechanisms and signaling pathways of autophagy. *Annu Rev Genet* **2009,** 43, 67-93.
42. Bandhakavi, S.; Xie, H.; O'Callaghan, B.; Sakurai, H.; Kim, D. H.; Griffin, T. J., Hsf1 activation inhibits rapamycin resistance and TOR signaling in yeast revealed by combined proteomic and genetic analysis. *PLoS One* **2008,** 3, (2), e1598.
43. Dephoure, N.; Gygi, S. P., Hyperplexing: a method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast. *Sci Signal* **2012,** 5, (217), rs2.
44. Tkach, J. M.; Yimit, A.; Lee, A. Y.; Riffle, M.; Costanzo, M.; Jaschob, D.; Hendry, J. A.; Ou, J.; Moffat, J.; Boone, C.; Davis, T. N.; Nislow, C.; Brown, G. W., Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. *Nat Cell Biol* **2012,** 14, (9), 966-76.
45. Hazelwood, L. A.; Daran, J. M.; van Maris, A. J.; Pronk, J. T.; Dickinson, J. R., The Ehrlich pathway for fusel alcohol production: a century of research on Saccharomyces cerevisiae metabolism. *Appl Environ Microbiol* **2008,** 74, (8), 2259-66.
46. Gibson, B. R.; Lawrence, S. J.; Leclaire, J. P.; Powell, C. D.; Smart, K. A., Yeast responses to stresses associated with industrial brewery handling. *FEMS Microbiol Rev* **2007,** 31, (5), 535-69.
47. Messenguy, F.; Dubois, E., Participation of transcriptional and post-transcriptional regulatory mechanisms in the control of arginine metabolism in yeast. *Mol Gen Genet* **1983,** 189, (1), 148-56.

48. Delbecq, P.; Werner, M.; Feller, A.; Filipkowski, R. K.; Messenguy, F.; Pierard, A., A segment of mRNA encoding the leader peptide of the CPA1 gene confers repression by arginine on a heterologous yeast gene transcript. *Mol Cell Biol* **1994,** 14, (4), 2378-90.
49. Lyutova, E. M.; Kasakov, A. S.; Gurvits, B. Y., Effects of arginine on kinetics of protein aggregation studied by dynamic laser light scattering and tubidimetry techniques. *Biotechnol Prog* **2007,** 23, (6), 1411-6.
50. Nishimura, A.; Kotani, T.; Sasano, Y.; Takagi, H., An antioxidative mechanism mediated by the yeast N-acetyltransferase Mpr1: oxidative stress-induced arginine synthesis and its physiological role. FEMS Yeast Res 2010, 10, (6), 687-98.
51. Domitrovic, T.; Palhano, F. L.; Barja-Fidalgo, C.; DeFreitas, M.; Orlando, M. T.; Fernandes, P. M., Role of nitric oxide in the response of Saccharomyces cerevisiae cells to heat shock and high hydrostatic pressure. *FEMS Yeast Res* **2003,** 3, (4), 341-6.
52. Davidson, J. F.; Whyte, B.; Bissinger, P. H.; Schiestl, R. H., Oxidative stress is involved in heat-induced cell death in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **1996,** 93, (10), 5116-21.
53. Moraitis, C.; Curran, B. P., Reactive oxygen species may influence the heat shock response and stress tolerance in the yeast Saccharomyces cerevisiae. *Yeast* **2004,** 21, (4), 313-23.
54. Nishimura, A.; Kawahara, N.; Takagi, H., The flavoprotein Tah18-dependent NO synthesis confers high-temperature stress tolerance on yeast cells. *Biochem Biophys Res Commun* **2013,** 430, (1), 137-43.
55. Gotz, R.; Gnann, A.; Zimmermann, F. K., Deletion of the carbonic anhydrase-like gene NCE103 of the yeast Saccharomyces cerevisiae causes an oxygen-sensitive growth defect. *Yeast* **1999,** 15, (10A), 855-64.
56. Cottier, F.; Raymond, M.; Kurzai, O.; Bolstad, M.; Leewattanapasuk, W.; Jimenez-Lopez, C.; Lorenz, M. C.; Sanglard, D.; Vachova, L.; Pavelka, N.; Palkova, Z.; Muhlschlegel, F. A., The bZIP transcription factor Rca1p is a central regulator of a novel CO(2) sensing pathway in yeast. *PLoS Pathog* **2012,** 8, (1), e1002485.
57. Lee, D. D.; Seung, H. S., Learning the parts of objects by non-negative matrix factorization. *Nature* **1999,** 401, (6755), 788-91.
58. Belle, A.; Tanay, A.; Bitincka, L.; Shamir, R.; O'Shea, E. K., Quantification of protein half-lives in the budding yeast proteome. *Proc Natl Acad Sci U S A* **2006,** 103, (35), 13004-9.
59. Wei, M.; Fabrizio, P.; Madia, F.; Hu, J.; Ge, H.; Li, L. M.; Longo, V. D., Tor1/Sch9-regulated carbon source substitution is as effective as calorie restriction in life span extension. *PLoS Genet* **2009,** 5, (5), e1000467.

Concluding Remarks

My first introduction to mass spectrometry based proteomics was with Dr. Gary Nelsestuen at the University of Minnesota. In his lab, we focused on clinical applications of mass spectrometry based proteomics. The goal of this research was to identify stable biomarkers in protein profiles of human plasma, which correlated with disease states using MALDI-TOF mass spectrometry. For example, we were able to identify and characterize a functional polymorphism of apolipoprotein C1, which was found in persons of Native American and Native Mexican ancestry; this variant was correlated with obesity and diabetes in those populations. On a larger scale, we identified several protein ratios and protein glycosylation states which correlated with various human diseases, including obesity, recovery from bariatric surgery, and bone marrow transplantation mortality. From this work, I drew several conclusions regarding the field of mass spectrometry based proteomics.

First, I realized how rapidly and sensitively data could be acquired through mass spectrometry techniques. Second, I was impressed by the wealth of data that is generated by such techniques and the requirement for robust bioinformatics software; indeed with the limited software availability, data analysis may require several orders of magnitude more time than the actual data acquisition. Third, I became aware of how reproducible mass spectrometry could be. Finally, through this work, I observed how directly applicable mass spectrometry based technology is to biology, particularly in the analysis of human disease; undeniably, the use of mass spectrometry is becoming more relevant directly in a clinical setting. This experience also introduced me to an important aspect of mass spectrometry, quantification using stable isotopes. In Dr. Nelsestuen's lab, for example, internal standards of deuterated plasma samples were routinely used for quantification.

With these experiences in mind, when I began my doctoral studies at Harvard Medical School, I gravitated toward labs which were interested in large-scale biology. Ultimately, after rotating through Dr. Gygi's lab, I became set on continuing my study in the field of proteomics. His lab had established robust methods for peptide identification (including the use of the reverse database strategy for false discovery rate estimation), SILAC quantification and the identification of phosphorylation sites. Despite many successes in the field of proteomics, several questions and technical considerations remained to be answered.

The level of proteomic coverage for both peptide and phosphopeptide identification was of constant interest to the field, and it was unclear what combinations of methods would facilitate the deepest proteome coverage. Another technical issue which existed when I joined the lab was the lack of a robust method for performing large scale quantitative proteomics in animal tissues. Regarding quantification, it was still unclear how robust mass spectrometry based quantification was in general, in terms of accuracy and reproducibility. Along with methodological concerns, bioinformatics analysis, including phosphorylation site assignment and quantification, and the effects of protein level data filtering were still ambiguous. Throughout my doctoral work, as technology improvements were introduced, the need to evaluate and demonstrate appropriate applications of these technologies was required. The biological relevance of data generated by these large scale endeavors is generally difficult to assess, and thus efforts to highlight relevant biology contained in large data sets are useful. Many of these issues were addressed throughout my dissertation.

Chapter two contained an analysis of phosphoproteomic depth through a comparison of IMAC and TiO_2 enrichment methods. The analysis demonstrated that greater than previously observed proteomic depth could be achieved (it was at the time the deepest phosphorylation analysis in yeast) by combining multiple technologies simultaneously. In addition, this project represented one of the earlier applications of the LTQ-Orbitrap for large scale phosphorylation analysis, helping to demonstrate that

the LTQ-Orbitrap could perform in a similar manner to the LTQ-FT-ICR. The LTQ-FT-ICR required constant predictive maintenance, and thus Orbitrap analysis is preferred. From a biological standpoint, many of the sites identified in chapter two were of relevance to the scientific community; many of these phosphorylation sites have been cited as the basis of additional research, or as a confirmation of identified and biologically relevant phosphorylation. In addition this study elaborated upon a fundamental goal of biological science, understanding results in an evolutionary context.

Chapter three addressed the question of finding a successful means for tissue based peptide quantification which was applicable on a large scale. In particular, the goal of this work was to obtain a robust strategy for quantitative phosphoproteomic analyses. Though peptide dimethylation has been used for years and had also been applied to mass spectrometry, we found published protocols to contain a vital flaw, improper reaction pH, which hindered their applications on a large scale. Addressing this issue was required for its application as a successful quantification strategy, and such an application was demonstrated in the comparison of fasted and re-fed mouse liver phosphorylation. Of relevance, a bioinformatic means for reducing isotopically labeled phosphopeptides down to localized and quantified sites was presented. Although such a consideration may seem trivial, this reduction complicated previous large-scale quantitative phosphorylation analyses, where much of it was performed manually. This chapter also examined proteomic depth through identifying the need for stringent protein level filtering in phosphoproteomics analyses. This study revealed a fundamental issue with MS^1 based quantification (linked identification and quantification) and offered solutions from which a more in depth analysis could be based. Chapter four offered additional solutions through the use of MS^2 based quantification.

Chapter four remarked on virtually all of the discussed questions in the field of mass spectrometry based proteomics. With the improvement in HCD fragmentation on the Orbitrap-Velos, for the first time MS^2 based quantification was applicable on a large scale. Previous MS^2 based

quantification studies generally relied on quadrupole-time-of-flight instruments, which powerful in their own right, could not match the proteomic depth of Orbitrap-Velos hybrid instrumentation. The presentation of HPRP vs. SCX chromatography for use in mass spectrometry sample preparation demonstrated that simple methodological improvements could greatly increase both analytical depth and robustness. In addition, the evaluation of the multinotch strategy for TMT quantification was demonstrated in a real biological system, an important consideration as this method is poised to become the preferred means of TMT quantification in the field mass spectrometry. With the inclusion of multiple labels, for the first time, replicates were able to be analyzed on a large scale, which demonstrated the reproducibility of quantification by TMT. This point is important, as mass spectrometry in often viewed in the field of biology as only a semi quantitative technique; in contrast, it was demonstrated in this chapter that accurate and reproducible quantification of thousands of proteins, from many samples simultaneously, was possible. Indeed, replicates in large scale analyses have become the norm, and greatly aid in downstream analyses. The bioinformatic requirements of these data sets were also addressed and the use of such systems level analytical tools as PCA and NMF were presented. Many of the tools which were previously only available to the genomics community are now applicable in the field of proteomics. Thus we can build upon the advancements that have already been made in the analysis of genomics data sets and apply it to our data. This chapter provides a solid foundation for future experimentation using quantitative multiplexing.

Taken together, the chapters of this dissertation demonstrate many successful strategies for the acquisition and analysis of proteome wide data sets. With the successful application of TMT in particular, we have the ability to perform true systems level analyses. Although technology will continue to improve over the coming years, I believe we are finally at the point where small scale techniques such as western blotting are no longer the gold standard. The perceived issues with mass spectrometry based proteomics have been adequately addressed by many labs, and the time has come for its robust

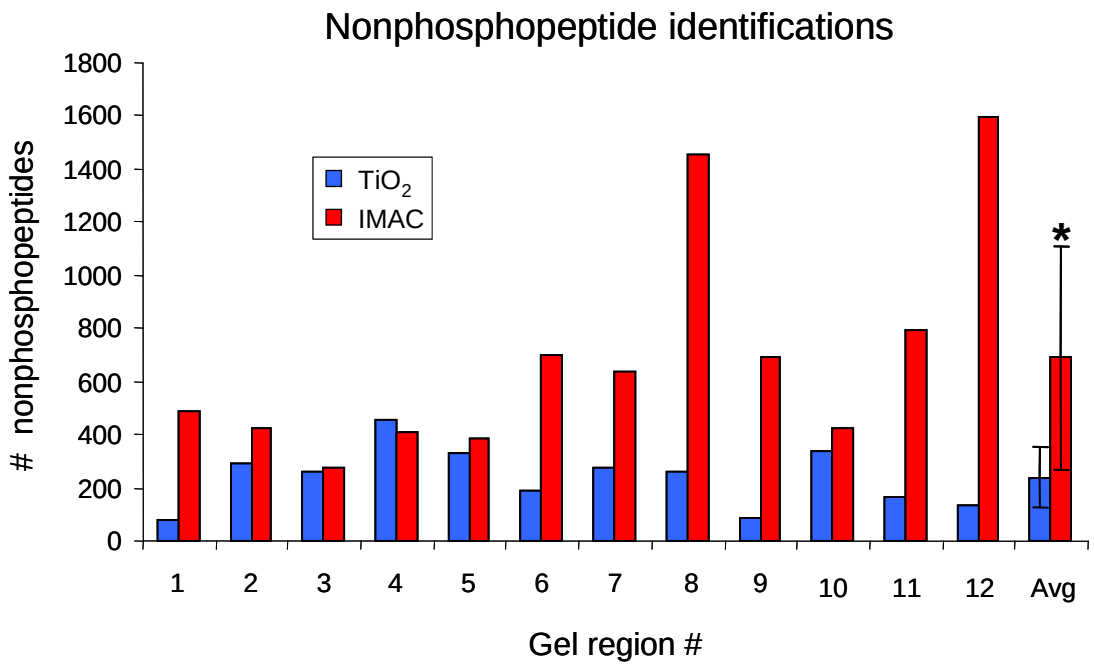
application in biology. We now have the ability to perform new analyses which were fundamentally unavailable with previous technologies. For example the combination of proteomics technology with other large scale multiplexed techniques, such as microarrays, may further aid in systems levels analyses by identifying the key points of gene product regulation, as they function in pathways or disease state. I do foresee, however, that improvements to computational tools and statistical analyses will always be of great importance. As such training in computer science and statistics should become as fundamental a component of graduate education as biochemistry or cell biology. To truly interpret large scale data in a meaningfully way, one must increasingly bridge the fields of biology and computation. With both a solid biological basis guiding the foundation of experimentation and the technological tools to answer big questions, the future of quantitative biology holds great promise for tackling the great mysterious in the natural sciences and human disease.

Appendix A

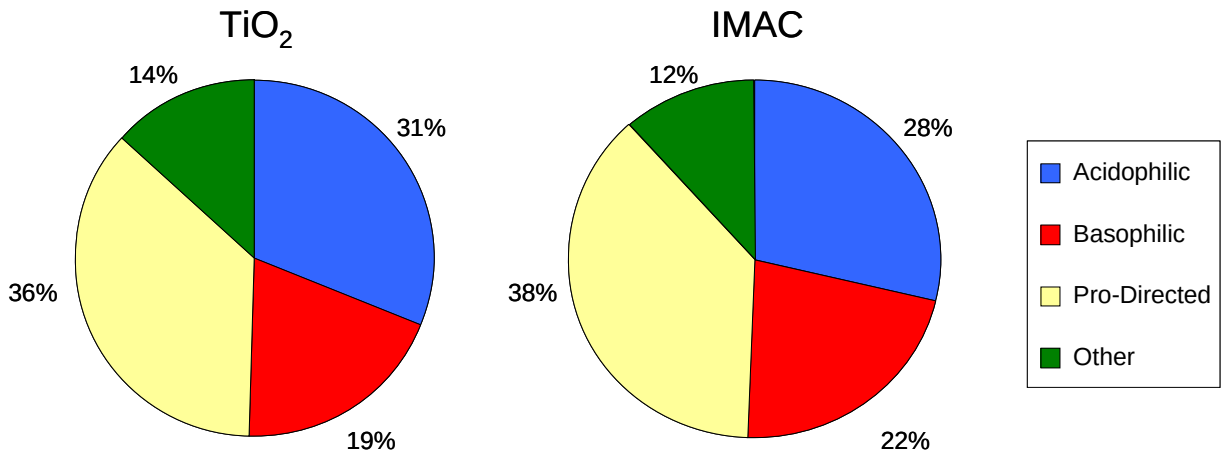
Supplemental Information

Chapter 2

A complete list of all phosphopeptides found in this study are presented in Supplemental Table 2.1. This table is attached electronically as an Excel file. A distribution of non-phosphorylated peptides by gel band for IMAC and TiO2 (Supplemental Figure 2.1). General motif classes for IMAC and TiO2 data sets (Supplemental Figure 2.2). A list of all motifs extracted from the data set using Motif-X (Supplemental Table 2.2). GO categories of the phosphoproteins in the data set for biological process and cellular localization (Supplemental Figure 2.3).



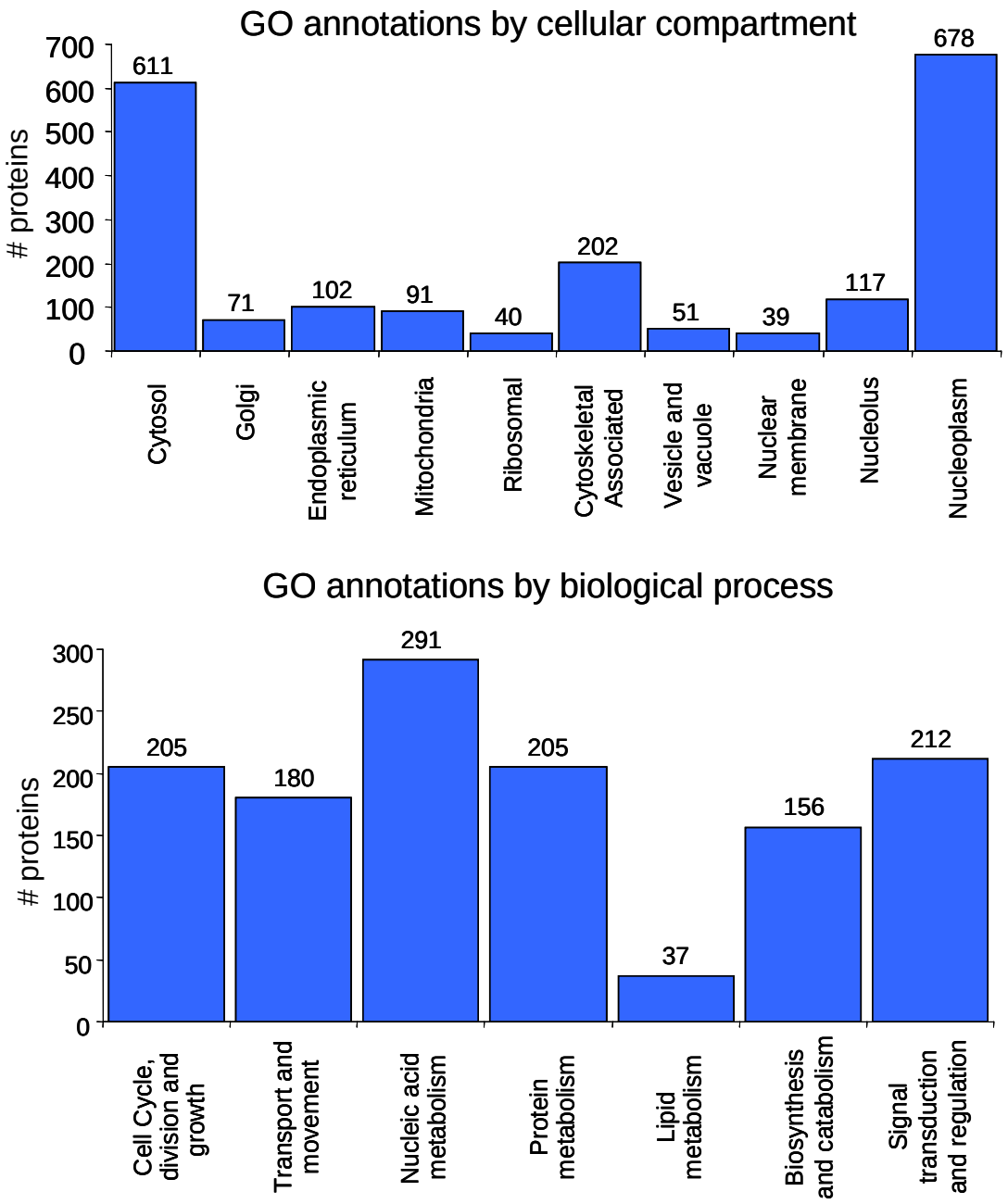
Supplemental Figure 2.1. Distribution of non-phosphorylated (contaminating) peptides across 12 gel bands. The distribution of non-phosphorylated peptides identified in the IMAC and TiO_2 enriched samples is shown. The IMAC enrichment method contained more non-phosphorylated peptides than the TiO_2 method. *P<0.007; Student's paired t-Test.



Supplemental Figure 2.2. General motif classes for IMAC and TiO_2 enriched peptides. General motif classes based on Villén et al [13]. The largest category was found to be general proline-directed phosphorylation, which might be expected based on high mitotic activity. No significant differences were observed between IMAC and TiO_2 enriched peptides. P = 1.0; Student's paired t-Test.

Supplemental Table 2.2. Analysis of singly phosphorylated motifs using Motif-X. Lower case S, T or Y indicate the phosphorylated residues and a "." represents any amino acid. The minimum significance was set at 10^{-6} for serine and threonine, and 10^{-4} for tyrosine. The motif score gives the significance of the motif and is equal to the sum of the negative log of the binomial probabilities used to calculate the motif (significance = $10^{-[\text{motif score}]}$). The fold increase of the motif is calculated by dividing the number of occurrences in the data set compared to the data set size by the number of occurrences in the whole proteome compared to the proteome size (of all s/t/y residues extended by 6aa of each side).

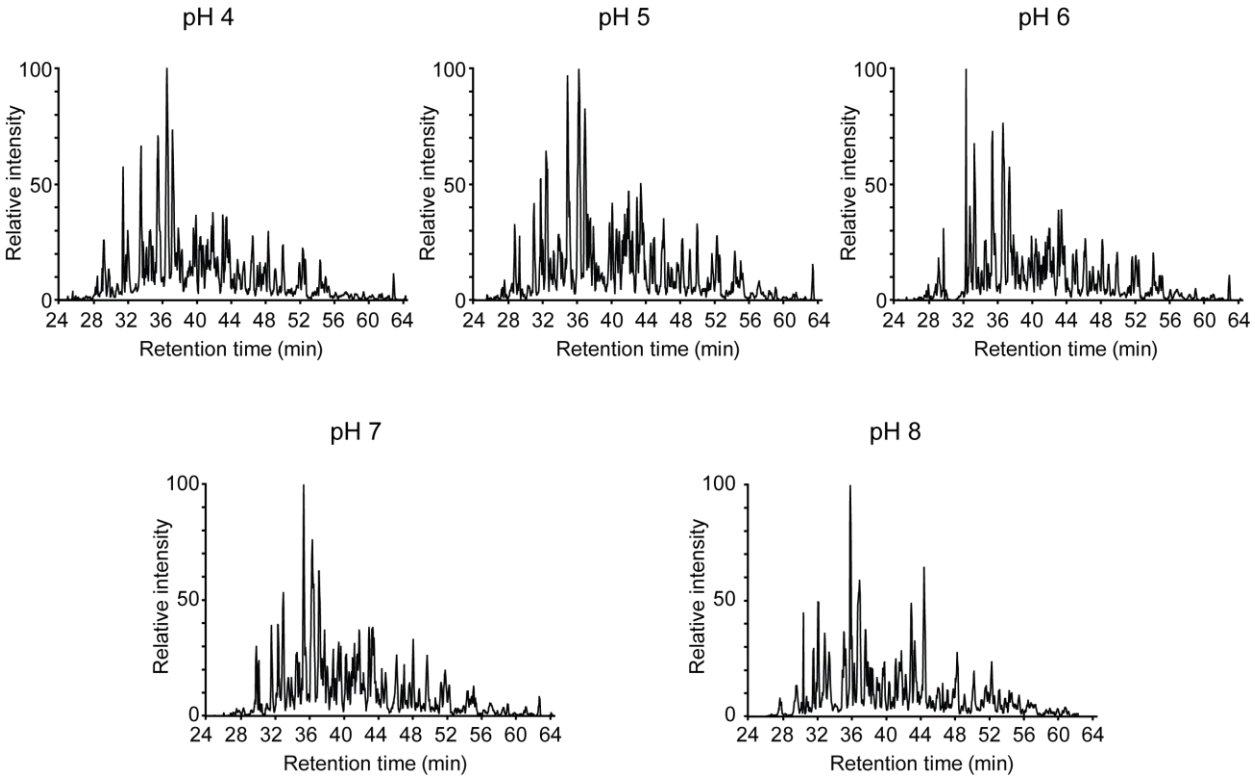
#	Motif	Motif Score	Foreground Matches	Foreground Size	Background Matches	Background Size	Fold Increase
1	...R..sP.....	26.71	67	1508	516	211881	18.24
2	......sPK....	22.23	56	1441	569	211365	14.44
3	......sPR....	22.23	40	1385	370	210796	16.45
4	......sP.....	16	376	1345	8765	210426	6.71
5	......sD.E...	32	73	969	935	201661	16.25
6	...R..s.....	16	202	896	9487	200726	4.77
7	......s.ED...	24.45	32	694	826	191239	10.68
8	......sE.E...	32	50	662	1054	190413	13.64
9	......sD.D...	30.81	34	612	615	189359	17.11
10	......s.D....	16	89	578	9800	188744	2.97
11	......s.E....	16	87	489	10039	178944	3.17
12	...K..s.....	12.09	67	402	10782	168905	2.61
13	...D..s.....	6.45	45	335	9412	158123	2.26
14	...R..tP.....	20.12	13	175	316	122651	28.83
15	......tPP....	20.03	12	162	319	122335	28.41
16	......tP.....	16	57	150	6384	122016	7.26
17	......tG.....	6.54	20	93	6664	115632	3.73
18	......y...R.Y	12.95	7	28	164	77455	118.07



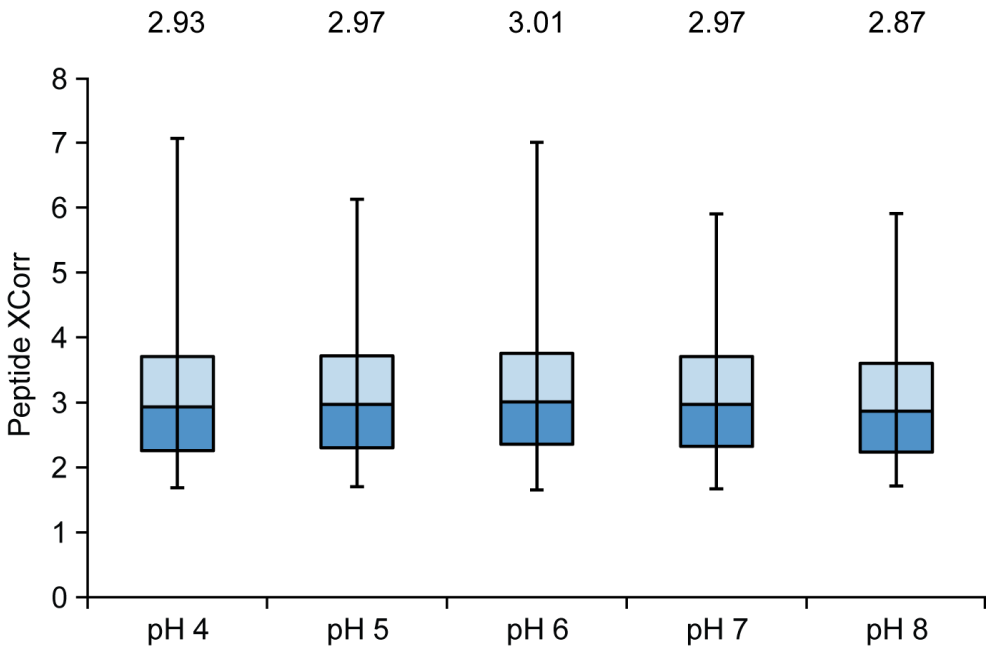
Supplemental Figure 2.3. Gene ontology classification of identified phosphoproteins. The biological processes and cellular localizations of the identified phosphoproteins were annotated with GO categories using the GoMiner program[29].

Chapter 3

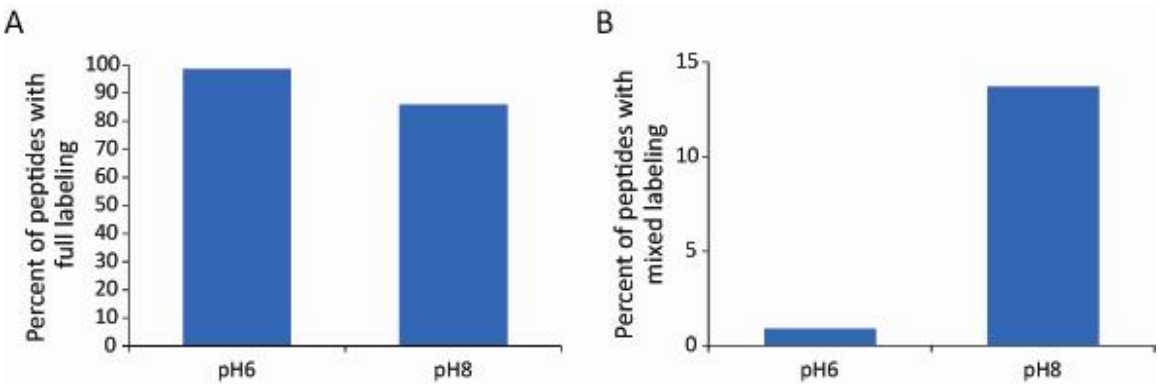
All phosphopeptides and phosphorylation sites identified in this study are presented in Supplemental Table 3.1, along with relevant properties (XCorr, heavy to light ratio, etc.). This table is attached electronically as an Excel file.



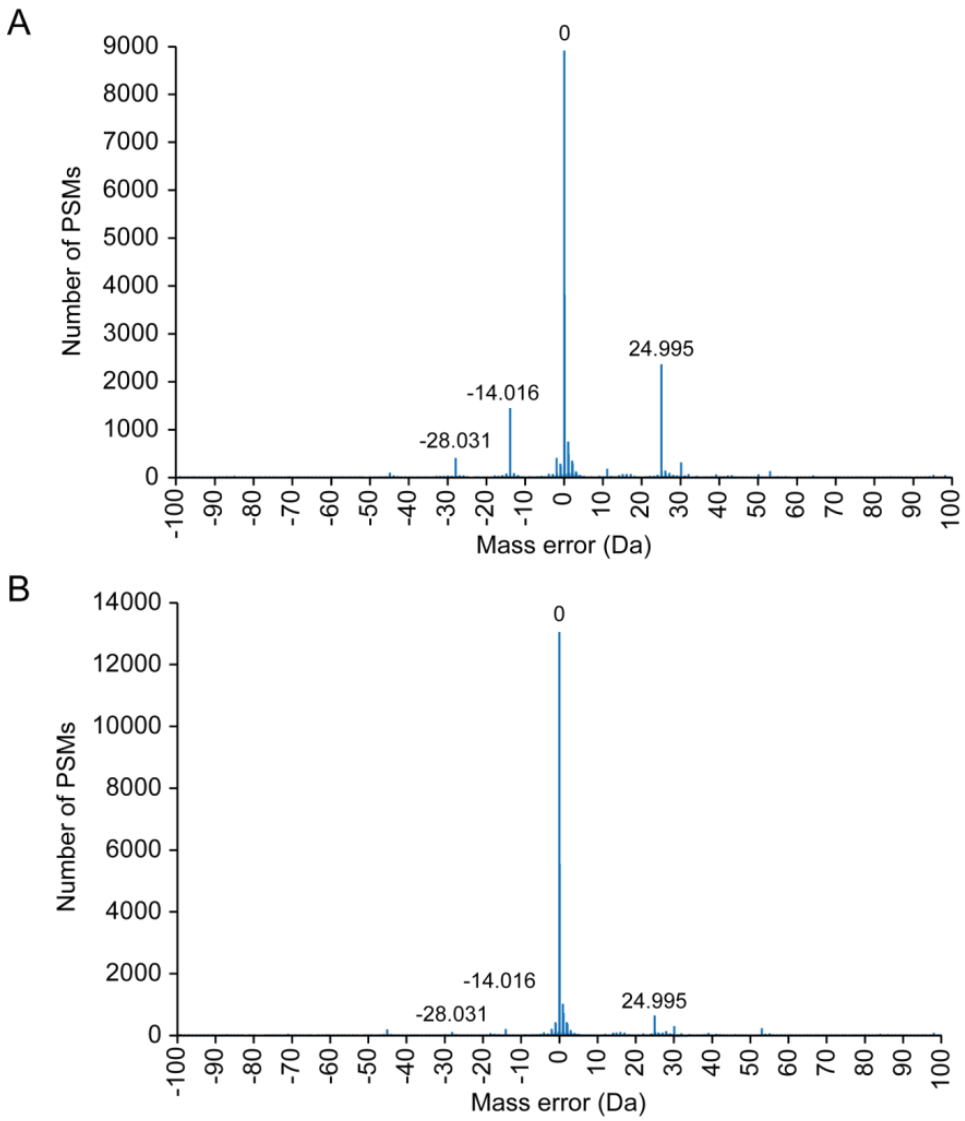
Supplemental Figure 3.1. The pH of the reductive dimethylation reaction does not affect the c18 reverse phased chromatographic elution of peptides. The base peak chromatograms (elution profile of the most intense ion in an MS^1 scan) of each analysis presented in Figure 3.2 on the mass spectrometer are plotted. The chromatograms were sufficiently similar in all cases to conclude that the pH of the dimethylation reaction does not affect chromatography of labeled peptides and that a liquid chromatography failure did not occur. Along with data in Figure 3.2 and Supplemental Figure 3.2, these data indicate that high pH reactions affect spectral matching and not the quality of the LC-MS/MS analysis. In all cases the maximum base peak was ~ 1×10^8.



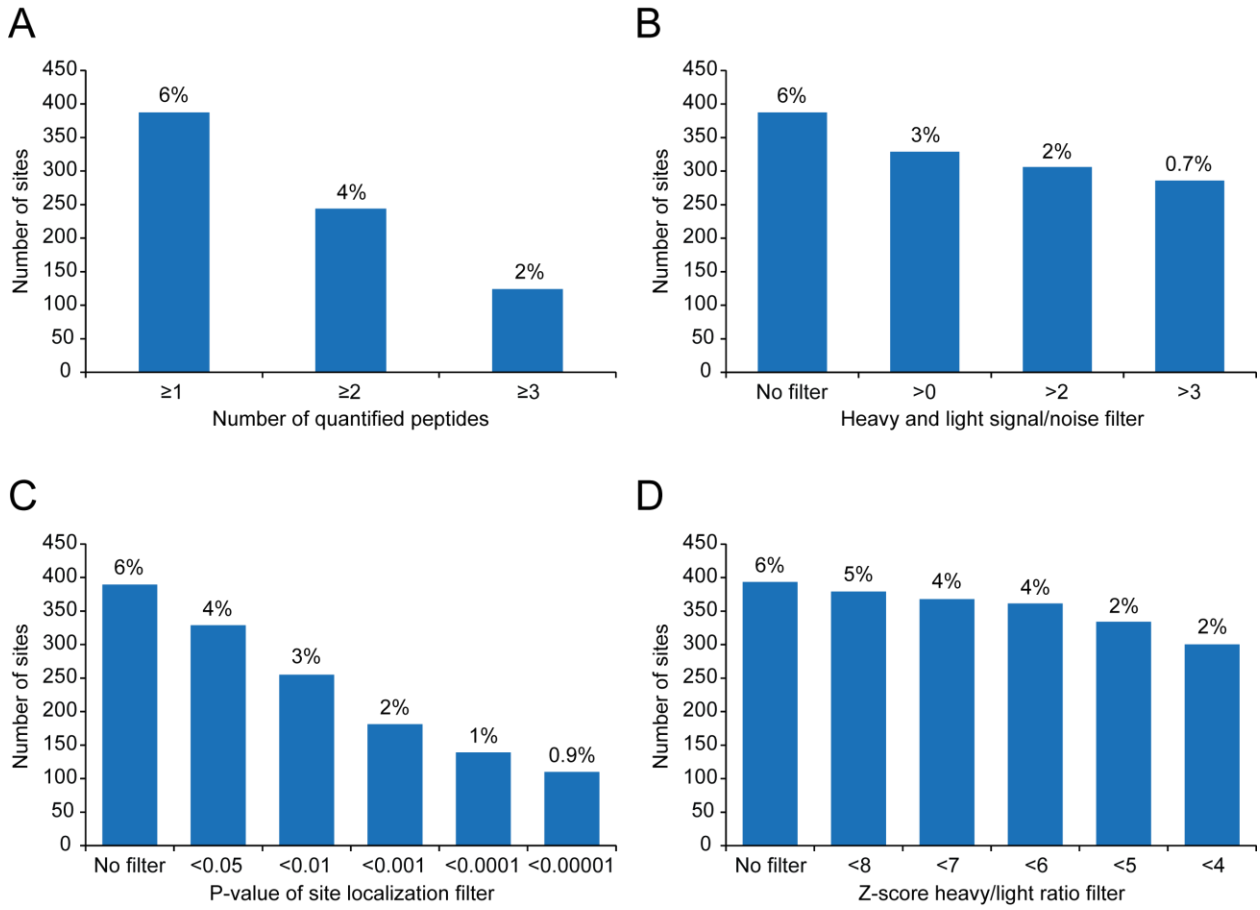
Supplemental Figure 3.2. Box plots of XCorr values suggest MS/MS data is of similar quality between various pH conditions. Box plots of XCorr values for matched peptides were created for each pH condition. Although the range of XCorr data was higher in the lower pH reaction conditions (pH 4-6), compared to higher pH conditions (pH 7 and 8), the median values and both quartiles (25^{th} and 75^{th} percentile) for all conditions fell within ~5 % of one another. The median XCorr value for each condition is listed above box. These data indicate that the MS/MS spectra were of similar quality for all reaction conditions, and differences in MS/MS quality were not responsible for the lower success rate in high pH conditions.



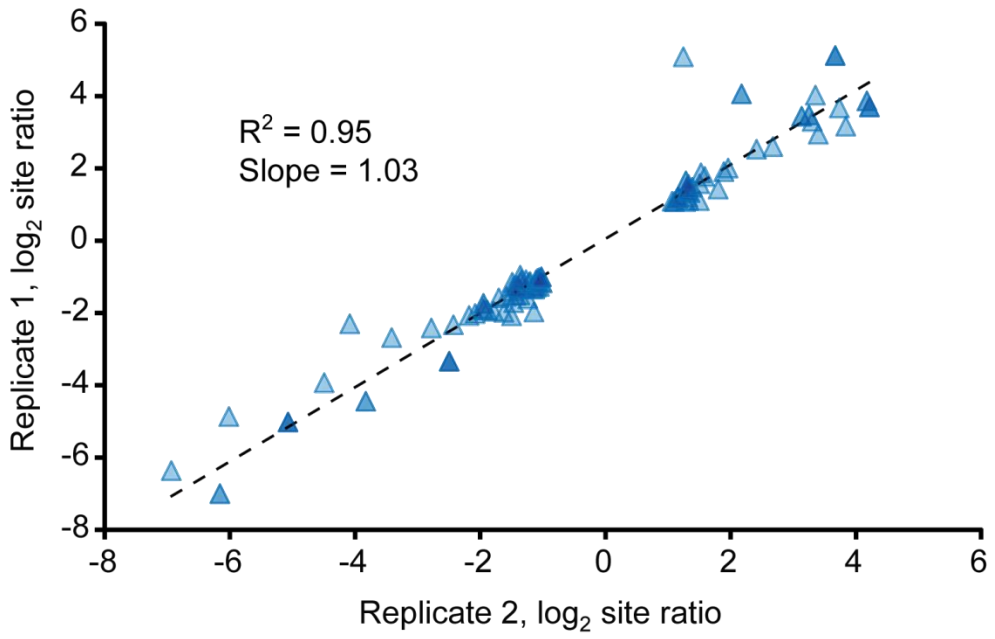
Supplemental Figure 3.3. The dimethylation reaction is quantitative at lower pH conditions (<6), but is only 85% efficient at higher pH conditions (pH 8). (A) The presence of non-fully labeled peptides at pH 6 (<1%) is within the tolerance of the FDR (1%). (B) The non-fully labeled peptides consist nearly exclusively of mixed labeled peptides, those with at least one methyl group addition, which are not fully dimethylated at all free amines (e.g. mono methylation at the peptide n-terminus and dimethylation at a lysine residue). The lack of complete labeling at pH8 contributes to the reduced peptide identification rate observed at this pH.



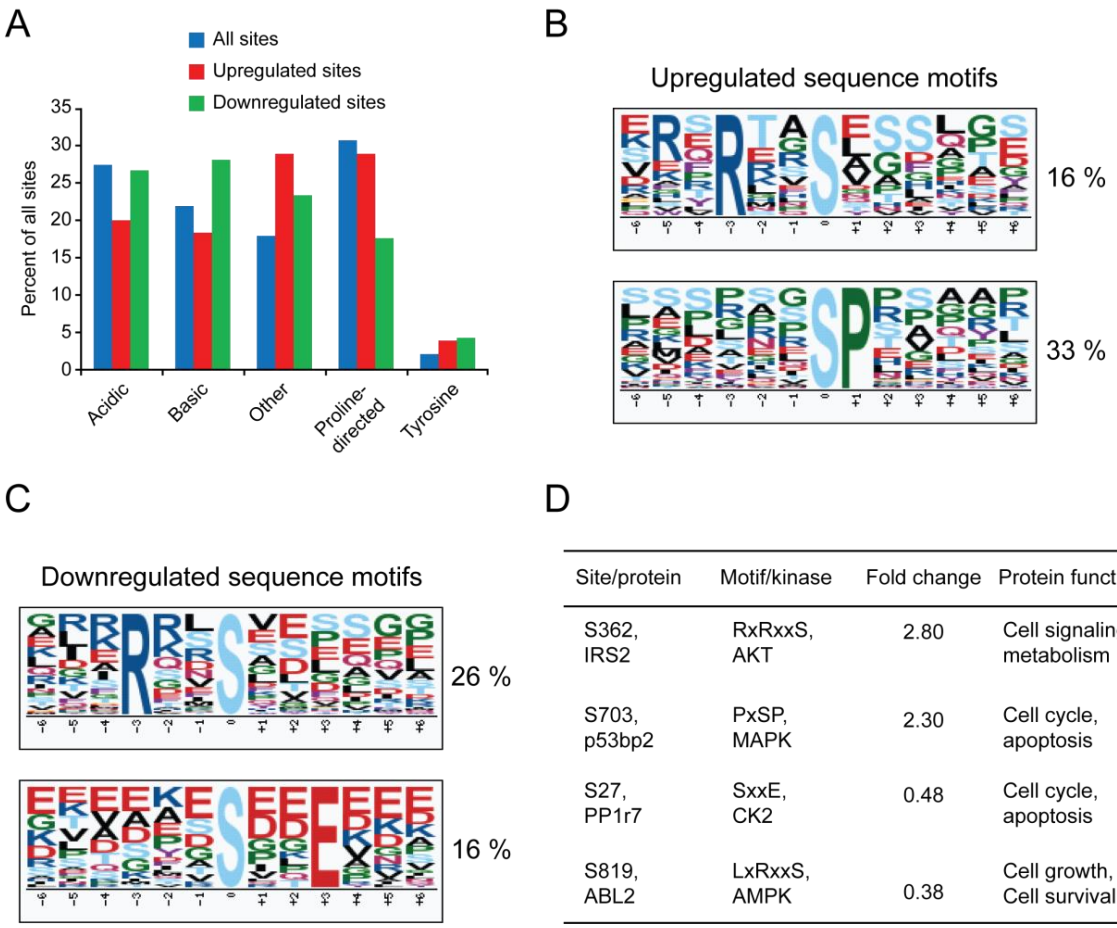
Supplemental Figure 3.4. Precursor ion mass error distributions (0.1 Da bins) for matched peptide spectra from dimethylation reactions performed at pH 8 (A) and 5.5 (B). Peptides were reacted with the light (non-deuterium containing) versions of formaldehyde and sodium cyanoborohydride only. Spectra were searched with a parent ion tolerance of 100 Da, static modifications of 57.0215 on cysteine (carbamidomethylation), 28.0313 on lysine and the peptide N-terminus (dimethylation), and the dynamic modification of 15.9949 on methionine (oxidation) was allowed. The data were filtered to a 1% FDR using linear discriminant analysis lacking the mass accuracy parameter. More spectra matched peptides at the correct mass (0 Da mass error bin) for the pH 5.5 reaction as compared to the pH 8 reaction. Noticeable peaks at -14 and -28 Daltons are present at pH 8, but not pH 5.5. These peaks likely consist of peptides lacking one and two dimethyl labels, respectively (combination of "mixed" labeled and possibly fully unlabeled peptides). These peaks support that labeling efficiency is reduced at pH 8. In addition a prominent peak at +25 Dalton is seen at pH 8 (and to a much lesser extent, pH 5.5). This mass (24.995) matches the addition of a cyano group to a peptide (while replacing a hydrogen atom). A list of known modifications can be found on the Unimod website (http://www.unimod.org/modifications_list.php?). This observation supports that side product formation is partially responsible for the reduced peptide identification rate seen high pH dimethylation reactions.



Supplemental Figure 3.5. Criteria for assessing confidence in regulated phosphorylation (two-fold change) sites. False positive hits tended to cluster in sites that changed by larger ratios. These sites may require additional criteria to judge the quality of their quantification. Filter criteria are listed on the x-axis, and the number of sites which passed each criterion is presented in columns on the y-axis. The estimated false discovery rates for each filter criterion are listed above each column. Requiring multiple peptides for the quantification of a given phosphorylation site allowed for increased confidence in the data set, albeit at a substantial loss to identifications (A). Signal-to-noise (S/N) filters effectively reduced the FDR without adversely affecting identifications (as compared to other criteria, B). As with the number of peptides used in quantification, site localization probability filtered the false discovery rate at a substantial cost to identifications (C). By removing outliers in the regulated data set (those with heavy to light ratios ≥5 standard deviations from the mean, Z-score) the FDR was effectively controlled in a similar manner to S/N (D). In all cases, filters were not intended to be hard cutoffs; rather, the listed criteria should help guide confidence in a given quantification event. As data set size increases, it may be possible to apply cutoff values for each criterion or filter the data through linear discriminant analysis, in a manner that does not adversely affect the identification of regulated sites.



Supplemental Figure 3.6. Technical replicates produce consistent phosphorylation site ratios for regulated phosphorylation sites. Only sites with a two-fold change (+/- 1 log_2 unit) are shown. The vast majority of sites that changed by two-fold were reproducibly quantified among technical replicates. The correlation between replicates was improved compared to Figure 3.5D (all ratios plotted by replicate). Since the majority of the observed sites changed with a ratio close to 1, small changes in the absolute ratio value lead to large changes in the percent difference between replicates. These small changes compounded over a large number of data points, negatively affecting the correlation of the two replicates. When these data points were removed, both the slope and the coefficient of determination approached 1.



Site/protein	Motif/kinase	Fold change	Protein function
S362, IRS2	RxRxxS, AKT	2.80	Cell signaling, metabolism
S703, p53bp2	PxSP, MAPK	2.30	Cell cycle, apoptosis
S27, PP1r7	SxxE, CK2	0.48	Cell cycle, apoptosis
S819, ABL2	LxRxxS, AMPK	0.38	Cell growth, Cell survival

Supplemental Figure 3.7. Upregulated and downregulated phosphorylation sites constitute similar motifs at different frequencies. Upregulated site were of greater abundance in the re-fed mice, whereas downregulated site were of greater abundance in the fasted mice. The distribution of general motif classes in the whole data set was compared to the general motif classes in the upregulated and downregulated data sets (A). The upregulated sites were deficient in acidic and basic motifs compared to all identified sites, whereas the downregulated sites were deficient in proline directed phosphorylation. The downregulated sites were enriched for basic motifs. Both the upregulated and downregulated datasets were enriched for uncharacterized motifs. Sequence specific web logos of extracted motifs from the motif-x program[24] for the upregulated and downregulated data sets are displayed with their frequencies (B and C).Though the upregulated data set was deficient in basic phosphorylation (A), the sequence specific motifs for PKA (RxxS) and AKT (RxRxxS) were often observed, along with many proline directed motifs (sP, B). More so than in the upregulated data set, the basic AKT/PKA and additionally AMPK (LxRxxS) type motifs were frequently downregulated (C). In addition the casein kinase II motif (SxxE) was extracted as a significant motif in the downregulated dataset. Based on these data, some of the observed changes cannot be simply explained by a change in kinase activity, but perhaps also by a change in substrate specificity of active kinases. Highlighted regulated sites which contain the extracted motifs, and an indication of which kinase may be responsible for its phosphorylating are indicated (D). IRS2 is a well-known mediator of insulin signaling. p53bp2 binds and affects the ability of p53 to bind DNA, and is involved in apoptosis[1]. The protein PP1r7 has been shown to inhibit protein phosphatase 1[2], and may affect cell cycle progression. Finally ABL2 is tyrosine kinase involved in a number of important biological processes including cell growth and survival. The enrichment of the discussed motifs is consistent with the role the related kinases may be playing in the context of energy metabolism and growth.

References

1. Samuels-Lev, Y.; O'Connor, D. J.; Bergamaschi, D.; Trigiante, G.; Hsieh, J. K.; Zhong, S.; Campargue, I.; Naumovski, L.; Crook, T.; Lu, X., ASPP proteins specifically stimulate the apoptotic function of p53. *Mol Cell* **2001,** 8, (4), 781-94.
2. Dinischiotu, A.; Beullens, M.; Stalmans, W.; Bollen, M., Identification of sds22 as an inhibitory subunit of protein phosphatase-1 in rat liver nuclei. *FEBS Lett* **1997,** 402, (2-3), 141-4.

Appendix B

CaMKIIbeta Signaling Pathway at the Centrosome Regulates Dendrite Patterning in the Brain

Attributions:

- This appendix contains work published as Puram, S. V., Kim, A. H., Ikeuchi, Y., Wilson-Grady, J. T., Merdes, A., Gygi, S. P., and Bonni, A., A CaMKIIbeta signaling pathway at the centrosome regulates dendrite patterning in the brain. Nat Neurosci 2011, 14, (8), 973-83

- J.T Wilson-Grady performed the LC-MS/MS analysis which identified the Ser 51 phosphorylation site, including database searching, site localization and manual validation, as well as contributed text to and edited the manuscript.

- S. V Puram, A. H. Kim and Y. Ikeuchi designed and performed *in vivo* experiments, biochemical assays and morphological analyses. S. V Puram and A. H. Kim prepared the manuscript.

- A. Merdes contributed molecular reagents.

- A. Bonni advised the project.

- S.P. Gygi provided instrumentation and computational tools for LC-MS/MS, and the relevant data analysis.

nature neuroscience

A CaMKIIβ signaling pathway at the centrosome regulates dendrite patterning in the brain

Sidharth V Puram[1–3], Albert H Kim[1,4], Yoshiho Ikeuchi[1], Joshua T Wilson-Grady[2,5], Andreas Merdes[6], Steven P Gygi[5] & Azad Bonni[1,2]

The protein kinase calcium/calmodulin-dependent kinase II (CaMKII) predominantly consists of the α and β isoforms in the brain. Although CaMKIIα functions have been elucidated, the isoform-specific catalytic functions of CaMKIIβ have remained unknown. Using knockdown analyses in primary rat neurons and in the rat cerebellar cortex *in vivo*, we report that CaMKIIβ operates at the centrosome in a CaMKIIα-independent manner to drive dendrite retraction and pruning. We also find that the targeting protein PCM1 (pericentriolar material 1) localizes CaMKIIβ to the centrosome. Finally, we uncover the E3 ubiquitin ligase Cdc20-APC (cell division cycle 20–anaphase promoting complex) as a centrosomal substrate of CaMKIIβ. CaMKIIβ phosphorylates Cdc20 at Ser51, which induces Cdc20 dispersion from the centrosome, thereby inhibiting centrosomal Cdc20-APC activity and triggering the transition from growth to retraction of dendrites. Our findings define a new, isoform-specific function for CaMKIIβ that regulates ubiquitin signaling at the centrosome and thereby orchestrates dendrite patterning, with important implications for neuronal connectivity in the brain.

The proper formation and morphogenesis of dendrites is fundamental to the establishment of neural circuits in the brain. After exit from the cell cycle and migration to the appropriate location, postmitotic neurons undergo carefully orchestrated steps in dendrite morphogenesis, including the generation and elaboration of extensive dendrite arbors followed by dendrite retraction and pruning[1,2]. These steps in dendrite patterning are necessary for the accurate formation of neuronal circuitry. Defects in dendrite patterning may result in severe neurodevelopmental disorders[3,4]. Although the molecular basis of dendrite growth has received scrutiny[5–8], the signaling mechanisms governing the transition from dendrite elaboration to pruning and retraction have remained poorly understood.

The calcium/calmodulin-dependent protein kinases (CaMKs) represent a critical link between the external environment and cellular responses in neurons. CaMKII is one of the major CaMKs in the brain, representing 1–2% of total brain protein[9–11]. CaMKII exists as a holoenzyme composed of multiple subunits, which form a complex through their association domains[12–15]. Brain CaMKII predominantly consists of the α and β isoforms, which form heteromeric or homomeric complexes[16–18]. Previous studies have focused on the functions of CaMKIIα[19–22]. However, a specific role for CaMKIIβ as a protein kinase in the mammalian brain remained unknown.

In this study, we report that CaMKIIβ has a distinct and specific function in the regulation of dendrite patterning in the mammalian brain. We identify a unique centrosomal targeting sequence (CTS) within the variable region of CaMKIIβ but not CaMKIIα. The CTS mediates the specific interaction of CaMKIIβ with the centrosomal targeting protein PCM1. Consequently, PCM1 localizes CaMKIIβ to the centrosome, where CaMKIIβ drives dendrite retraction and pruning independently of CaMKIIα. Notably, we uncover the ubiquitin ligase Cdc20-APC as the centrosomal substrate of CaMKIIβ. CaMKIIβ phosphorylates the APC coactivator Cdc20 at Ser51 in neurons, which induces Cdc20 dispersion from the centrosome, thereby inhibiting centrosomal Cdc20-APC activity and triggering a switch from growth to retraction of dendrites. Our findings define a CaMKIIβ pathway that regulates ubiquitin signaling at the centrosome and thereby orchestrates dendrite patterning and hence the establishment of neuronal connectivity in the mammalian brain.

RESULTS

CaMKIIβ regulates dendrite patterning in mammalian neurons

To characterize the role of CaMKIIβ in neurons, we used granule neurons of the developing rat cerebellar cortex. Granule neurons are the most abundant neurons in the brain and follow a highly typified spatial and temporal program of differentiation characteristic of neurons in the central nervous system, providing a robust system for studies of neuronal morphogenesis and connectivity[23–25]. Using an antibody that specifically recognizes CaMKIIβ, we first confirmed that the distinct isoforms of CaMKIIβ (β, β′, βε and β′ε) are expressed in primary granule neurons, and protein levels increased with maturation (Fig. 1a).

[1]Department of Pathology, Harvard Medical School, Boston, Massachusetts, USA. [2]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA. [3]MD-PhD Program, Harvard Medical School, Boston, Massachusetts, USA. [4]Department of Neurosurgery, Brigham and Women's Hospital, Children's Hospital, Boston, Massachusetts, USA. [5]Department of Cell Biology, Harvard Medical School, Boston, Massachusetts, USA. [6]Centre de Recherche en Pharmacologie-Santé, Unité Mixte de Recherche 2587, Centre National de la Recherche Scientifique–Pierre Fabre, Toulouse, France. Correspondence should be addressed to A.B. (azad_bonni@hms.harvard.edu).

Received 11 March; accepted 28 April; published online 3 July 2011; doi:10.1038/nn.2857

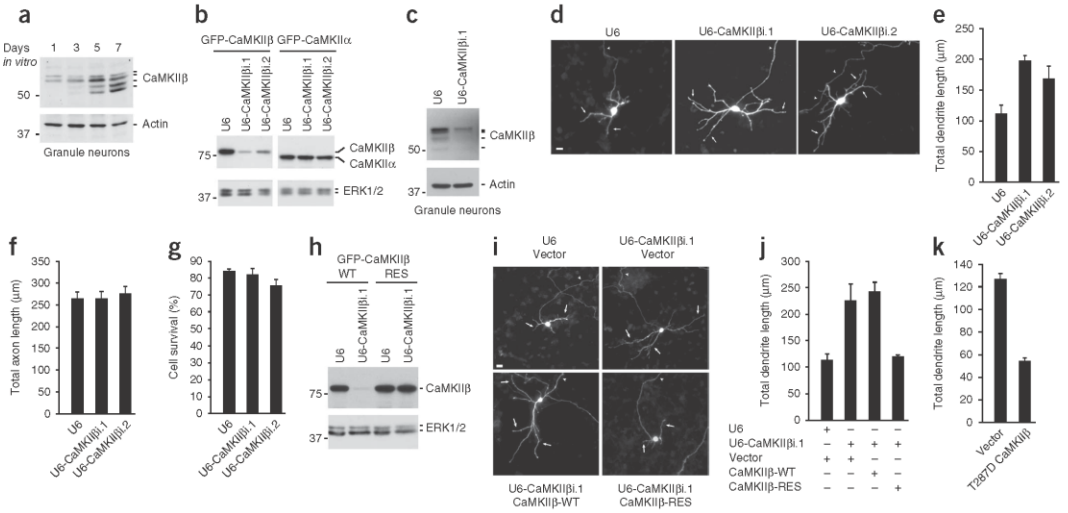


Figure 1 CaMKIIβ restricts the elaboration of dendrites. (**a**) Immunoblot of granule neuron lysates. Molecular weight in kDa is indicated in all blots. Full-length blots for all immunoblot analyses are presented in **Supplementary Figure 11**. (**b**) Immunoblot of lysates from COS cells transfected with GFP-CaMKIIβ or GFP-CaMKIIα together with the CaMKIIβ RNAi or control U6 plasmid. ERK1/2, extracellular signal-regulated kinases 1 and 2. (**c**) Immunoblot of lysates from granule neurons electroporated with the CaMKIIβ RNAi or control U6 plasmid. (**d**) Granule neurons transfected with a CaMKIIβ RNAi or control U6 plasmid together with GFP were subjected to immunocytochemistry using the GFP antibody. In all images of neuronal morphology, arrows and arrowheads indicate dendrites and axons, respectively. Scale bar, 10 μm. (**e**) Total dendrite length for granule neurons treated as in **d**. Total dendrite length was significantly greater in CaMKIIβ knockdown neurons than in control U6-transfected neurons (analysis of variance (ANOVA), $P < 0.005$). In total, 270 neurons were measured. (**f**) Granule neurons were analyzed as in **d**. Total axon length was not significantly different between CaMKIIβ knockdown neurons and control U6-transfected neurons. (**g**) Granule neurons were transfected with a CaMKIIβ RNAi or control U6 plasmid and analyzed for cell survival. Cell survival was not significantly different between CaMKIIβ knockdown neurons and control U6-transfected neurons. (**h**) Immunoblot of lysates from COS cells transfected with GFP-CaMKIIβ-WT or GFP-CaMKIIβ-RES together with the CaMKIIβ RNAi or control U6 plasmid. (**i**) Granule neurons transfected with the CaMKIIβ RNAi or control U6 plasmid together with the expression plasmid encoding CaMKIIβ-WT, CaMKIIβ-RES or control vector and GFP, analyzed as in **d**. Scale bar, 10 μm. (**j**) Total dendrite length for granule neurons treated as in **i** was quantified. CaMKIIβ-RES, but not CaMKIIβ-WT, significantly reduced total dendrite length compared to control vector in the background of CaMKIIβ RNAi (ANOVA, $P < 0.005$). In total, 360 neurons were measured. (**k**) Granule neurons transfected with constitutively active T287D CaMKIIβ or control vector were analyzed as in **d**. Total dendrite length was significantly lower in T287D CaMKIIβ-expressing neurons than in control vector-transfected neurons (*t*-test, $P < 0.0005$). In total, 180 neurons were measured. Error bars, s.e.m.

To determine the function of CaMKIIβ in neurons, we used a plasmid-based method of RNA interference (RNAi) to acutely knockdown CaMKIIβ expression[26]. Expression of two short hairpin RNAs (shRNAs) targeting distinct regions of CaMKIIβ reduced the expression of exogenous CaMKIIβ but not CaMKIIα in COS simian kidney cells (**Fig. 1b**), and robustly decreased the expression of endogenous CaMKIIβ but not CaMKIIα in granule neurons as determined by immunoblotting and immunocytochemistry (**Fig. 1c** and **Supplementary Fig. 1a–c**). To assess the role of CaMKIIβ in neuronal morphogenesis, we induced the knockdown of CaMKIIβ in granule neurons and examined axons and dendrites, which are easily identified based on their morphology and immunocytochemical markers[27,28]. Notably, CaMKIIβ knockdown in granule neurons stimulated the elaboration of dendrites, which were characterized by longer primary dendrites and more secondary and tertiary dendrite branching (**Fig. 1d** and **Supplementary Fig. 1d**). Accordingly, CaMKIIβ knockdown increased total dendrite length by up to 76% in granule neurons (**Fig. 1e**). By contrast, CaMKIIβ RNAi had little or no effect on axon length (**Fig. 1f**). In other experiments, CaMKIIβ RNAi had little or no effect on cell survival in granule neurons (**Fig. 1g**). Together, these results suggest that CaMKIIβ knockdown specifically stimulates the elaboration of dendrites in granule neurons.

To confirm that the CaMKIIβ RNAi-induced dendrite phenotype was the result of specific knockdown of CaMKIIβ, we performed a rescue experiment. We generated an expression plasmid encoding CaMKIIβ that is resistant to RNAi (CaMKIIβ-RES) (**Fig. 1h**). Expression of CaMKIIβ-RES, but not CaMKIIβ encoded by wild-type cDNA (CaMKIIβ-WT), restored the typical appearance of dendrite arbors and reduced dendrite length and branching in the background of CaMKIIβ RNAi to that of control transfected neurons (**Fig. 1i,j** and **Supplementary Fig. 1e**). These data suggest that the CaMKIIβ RNAi-induced dendrite phenotype results from specific knockdown of CaMKIIβ rather than off-target effects of CaMKIIβ RNAi. In other experiments, expression of a constitutively active form of CaMKIIβ, in which the site of autophosphorylation Thr287 is replaced with the phosphomimetic residue aspartate (T287D CaMKIIβ), simplified dendrite arbors and profoundly reduced dendrite length and branching in granule neurons (**Fig. 1k** and data not shown). Thus, based on both inhibition of CaMKIIβ by rigorously controlled RNAi and gain-of-function analyses, we conclude that CaMKIIβ restricts the extent of dendrite arborization and growth in granule neurons.

We next determined whether the function of CaMKIIβ in dendrite patterning is generalizable in mammalian brain neurons. CaMKIIβ knockdown in hippocampal neurons substantially increased the number of secondary and tertiary dendrite branches and thereby increased dendrite length (**Supplementary Fig. 2a–d**). CaMKIIβ RNAi-induced dendrite elaboration in hippocampal

neurons was reversed by CaMKIIβ-RES but not CaMKIIβ-WT, indicating the specificity of the CaMKIIβ RNAi-induced phenotype (**Supplementary Fig. 2e**,**f**). CaMKIIβ knockdown also stimulated dendrite elaboration in primary cerebral cortical neurons (**Supplementary Fig. 3**). These data suggest that CaMKIIβ inhibits the growth and arborization of dendrites in diverse populations of mammalian brain neurons.

To determine the cellular basis of CaMKIIβ regulation of dendrite patterning, we characterized the temporal dynamics of CaMKIIβ function in dendrite morphogenesis. We subjected individual granule neurons expressing constitutively active T287D CaMKIIβ and control vector-transfected granule neurons to time-lapse analyses. In control vector-transfected granule neurons, dendrites and their branches dynamically elongated and retracted during 48 h of analysis (**Supplementary Fig. 4a**,**b**), with an overall cumulative increase in total dendrite length (**Supplementary Fig. 4c–e**). By contrast, T287D CaMKIIβ-expressing granule neurons rarely had periods of dendrite growth and almost exclusively retracted their dendrites (**Supplementary Fig. 4a**,**b**), which cumulatively reduced total dendrite length (**Supplementary Fig. 4c–e**). Analyses of individual dendrites yielded similar results, with individual dendrites in T287D CaMKIIβ-expressing neurons also retracting more often than individual dendrites in control granule neurons (**Supplementary Fig. 4f–h**). In a complementary line of experiments, time-lapse analyses of individual neurons and dendrites revealed reduced periods of dendrite retraction and increased periods of dendrite growth, with an overall cumulative increase in total dendrite length in CaMKIIβ knockdown neurons (**Supplementary Fig. 4i–q**). These data suggest that CaMKIIβ restricts the elaboration of dendrite arbors by triggering a switch from growth to active retraction of dendrites.

We next characterized the function of CaMKIIβ in regulating the minute dynamics of dendrite extension and retraction using live environment-controlled imaging analyses. Granule neurons expressing T287D CaMKIIβ or transfected with the control vector were imaged every 10 min for a 1 h period. In control neurons, dendrites had many extension events with few retraction events (**Supplementary Fig. 5a**,**b** and **Supplementary Video 1**). By contrast, T287D CaMKIIβ-expressing neurons had many retraction events and few extension events (**Supplementary Fig. 5a**,**b** and **Supplementary Video 1**). In a complementary line of experiments, live imaging analyses revealed a substantial increase in extension events and a substantial reduction in retraction events in neurons upon CaMKIIβ knockdown (**Supplementary Fig. 5c**,**d** and **Supplementary Video 2**).

CaMKIIβ drives dendrite retraction and pruning *in vivo*

Having identified a crucial function for CaMKIIβ in the control of dendrite patterning in primary neurons, we next determined the role of CaMKIIβ in dendrite morphogenesis in the intact developing cerebellar cortex. We first used rat organotypic cerebellar slices in which the architecture of the cerebellar tissue is preserved. Using a biolistics method of transfection, we induced CaMKIIβ knockdown in postnatal day 6 (P6) cerebellar slices[27]. CaMKIIβ knockdown robustly stimulated dendrite elaboration in internal granule layer (IGL) granule neurons (**Fig. 2a**), substantially increasing the number of secondary and tertiary dendrite branches as well as total dendrite length (**Fig. 2a**,**b** and **Supplementary Fig. 6a**).

We next assessed CaMKIIβ function in the cerebellar cortex in the organism using an *in vivo* RNAi approach[28]. We electroporated P3 rat pups with a CaMKIIβ RNAi plasmid that also expresses green fluorescent protein (U6-CaMKIIβi/CMV-GFP) or the corresponding

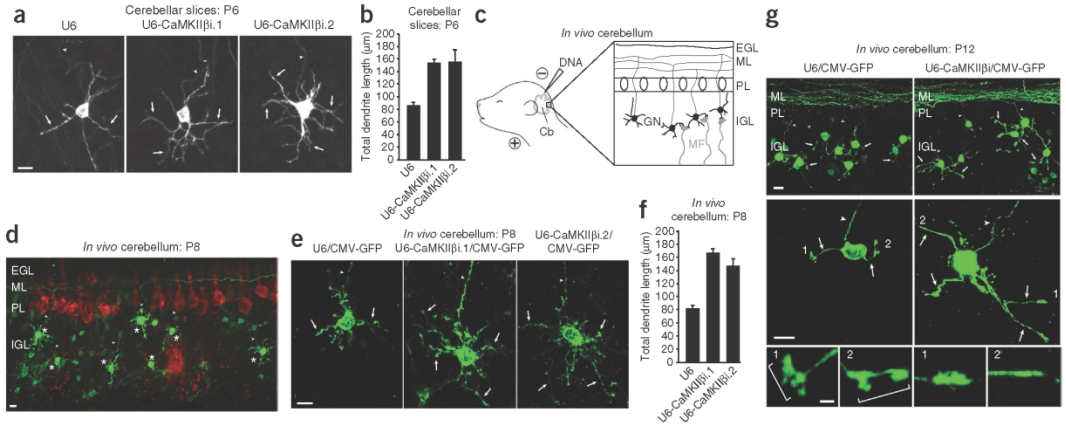


Figure 2 CaMKIIβ regulates dendrite patterning *in vivo*. (**a**) P6 rat cerebellar slices transfected by a biolistics method with the CaMKIIβ RNAi or control U6 plasmid together with GFP were subjected to immunohistochemistry using the GFP antibody. Scale bar, 10 μm. (**b**) CaMKIIβ knockdown neurons had significantly longer dendrites than control U6-transfected neurons (ANOVA, $P < 0.0001$). In total, 199 neurons were measured. (**c**) Schematic of *in vivo* electroporation approach. Cb, cerebellum; EGL, external granule layer; ML, molecular layer; PL, Purkinje cell layer; IGL, internal granule layer; GN, granule neurons; MF, mossy fiber. (**d**) Rat pups electroporated *in vivo* with a U6-CaMKIIβi/CMV-GFP RNAi or control U6/CMV-GFP plasmid were killed 5 d after electroporation and cerebella were subjected to immunohistochemistry using the GFP and calbindin antibodies. Representative control-transfected granule neurons are shown. Asterisks indicate cell somas. Scale bar, 10 μm. (**e**) Rat pups were electroporated as in **d**, and representative neurons for each condition are shown. Scale bar, 10 μm. (**f**) IGL granule neurons analyzed as in **d** were subjected to morphometric analysis. Total dendrite length was significantly greater in IGL granule neurons in CaMKIIβ knockdown rats than in control U6 rats (ANOVA, $P < 0.0001$). In total, 266 neurons were measured. (**g**) Rat pups electroporated *in vivo* with the U6-CaMKIIβi/CMV-GFP RNAi or control U6/CMV-GFP plasmid were killed 9 d after electroporation and cerebella were analyzed as in **d**. Top two panels, representative cerebellar sections from each condition. Scale bar, 10 μm. Middle two panels, representative IGL granule neurons for each condition. Scale bar, 10 μm. Bottom two panels, zoomed views of dendritic tips of individual neurons. Scale bar, 2.5 μm. Brackets identify dendritic claws. Error bars, s.e.m.

control RNAi plasmid (U6/CMV-GFP) (Fig. 2c). Five days after electroporation, rats were killed and cerebella were analyzed immunohistochemically using the GFP antibody (Fig. 2d). At P8, IGL granule neurons in CaMKIIβ knockdown rats had longer dendrites with more secondary and tertiary dendrite branching than IGL granule neurons in control rats (Fig. 2e). Morphometric analyses revealed substantially more secondary and tertiary dendrite branches and nearly twice the total dendrite length in IGL neurons in CaMKIIβ knockdown rats as in control rats (Fig. 2f and Supplementary Fig. 6b). CaMKIIβ knockdown had little or no effect on parallel fiber patterning or the number of parallel fibers associated with IGL granule neurons (Supplementary Fig. 6c and data not shown). Together, these data reveal a physiological, cell-autonomous function for CaMKIIβ in restricting the elaboration of dendrite arbors in the cerebellar cortex *in vivo*.

Because CaMKIIβ triggered active retraction of dendrites in primary neurons (Supplementary Figs. 4 and 5), we reasoned that CaMKIIβ might also function in pruning of dendrite arbors *in vivo*, which follows the phase of dendrite growth and arborization[1]. To test this possibility, we analyzed the effect of CaMKIIβ knockdown on dendrite morphogenesis in P12 rat pups *in vivo*. In control transfected rats, IGL granule neurons had a few short dendrites with simplified arbors (Fig. 2g and Supplementary Fig. 6d–g), characteristic of the mature stage of dendrite differentiation in granule neurons. In addition, IGL granule neurons harbored dendritic claws (Fig. 2g and Supplementary Fig. 6h), which house synapses with afferent mossy fiber terminals and Golgi neuron axons[1,2,29–32], providing further evidence of dendrite maturation in control P12 rats. In contrast to control rats, IGL granule neurons in CaMKIIβ knockdown rats had longer, more branched dendrite arbors (Fig. 2g), with greater total dendrite length (78.3 ± 3.0 μm in CaMKIIβ knockdown rats versus 43.8 ± 1.5 μm in control rats; *t*-test, $P < 0.0001$), primary dendrite number (3.53 ± 0.10 in CaMKIIβ knockdown rats versus 3.01 ± 0.08 in control rats; *t*-test, $P < 0.005$) and secondary and tertiary dendrite branch number (1.13 ± 0.11 in CaMKIIβ knockdown rats versus 0.63 ± 0.07 in control rats; *t*-test, $P < 0.005$) (Supplementary Fig. 6d–g), suggesting that CaMKIIβ knockdown impairs dendrite pruning and blocks the differentiation of dendrites at the stage of exuberant arbors. Consistent with these observations, IGL granule neuron dendrites in CaMKIIβ knockdown pups had a lower percentage of dendrites bearing claws (31.9 ± 2.2% in CaMKIIβ knockdown rats versus 58.9 ± 2.5% in control rats; *t*-test, $P < 0.0001$) (Fig. 2g and Supplementary Fig. 6h). Together, these results suggest that CaMKIIβ is essential for dendrite pruning at later stages of dendrite morphogenesis in the cerebellar cortex *in vivo*. Similar results were obtained in P10 cerebellar slices (Supplementary Fig. 6i,j), corroborating the conclusion that CaMKIIβ promotes dendrite pruning during later stages of dendrite development in the cerebellar cortex. Collectively, our findings suggest that CaMKIIβ regulates the patterning of dendrites throughout development *in vivo*.

CaMKIIβ functions independently of CaMKIIα at the centrosome

The identification of a role for CaMKIIβ in the control of dendrite patterning in the mammalian brain led us to determine the molecular basis of CaMKIIβ function in neurons. We exploited the ability of CaMKIIβ-RES to rescue the CaMKIIβ RNAi phenotype to perform structure–function analyses of CaMKIIβ in the regulation of dendrite morphogenesis. In contrast to CaMKIIβ-RES, a catalytically inactive mutant of CaMKIIβ-RES in which the ATP binding site was disrupted (K43R CaMKIIβ-RES) failed to restrict the elaboration of dendrites in the background of CaMKIIβ RNAi in granule neurons (Fig. 3a), suggesting that CaMKIIβ requires the catalytic activity of the kinase to control dendrite patterning.

The kinase domain in CaMKIIβ was required to inhibit growth and stimulate retraction of dendrites, and yet CaMKIIα, whose catalytic domain is very similar to CaMKIIβ's, promotes dendrite growth and arborization[27,33]. We therefore reasoned that CaMKIIβ but not CaMKIIα might be localized to a subcellular site that endows CaMKIIβ specifically with the ability to phosphorylate local substrates and restrict the elaboration of dendrites. CaMKIIβ harbors an F-actin binding domain (FABD) within the N-terminal portion of the variable region, but CaMKIIβ bound to F-actin seems to function independently of its catalytic domain as a scaffold protein that recruits CaMKIIα to F-actin[34–36]. Accordingly, deletion of the FABD (ΔFABD) had little or no effect on the ability of CaMKIIβ-RES to restrict dendrite elaboration in the background of CaMKIIβ RNAi in granule neurons (Fig. 3b,c). We next focused on the C-terminal segment of the variable region (CTRv), whose function is unknown. Notably, a CaMKIIβ-RES mutant in which the CTRv was deleted (CaMKIIβ-RESΔCTRv) failed to restrict dendrite elaboration in granule neurons in the background of CaMKIIβ RNAi (Fig. 3b,c). These results suggest that the CTRv is required for CaMKIIβ-regulation of dendrite patterning.

To characterize the function of the CTRv within CaMKIIβ, we assessed whether the CTRv localizes CaMKIIβ to a distinct subcellular site where CaMKIIβ controls dendrite patterning. A GFP-fusion protein of the CTRv appeared as a perinuclear punctate signal that colocalized with the centrosomal protein pericentrin (Fig. 3d). In immuno-electron microscopy analyses, a pool of endogenous CaMKIIβ was present in the pericentriolar region in granule neurons (Supplementary Fig. 7a). In complementary biochemical fractionation experiments using granule neuron lysates, a pool of endogenous CaMKIIβ fractionated with the centrosomal proteins γ-tubulin, Cdc20 and 14-3-3ε (Fig. 3e). Notably, endogenous CaMKIIα did not appear in the centrosomal fraction despite the presence of CaMKIIα in the whole cell lysate (Fig. 3e). Together, these data demonstrate that the CTRv represents a CTS that localizes CaMKIIβ to the centrosome in neurons.

If the principal function of the CTS is to localize CaMKIIβ to the centrosome, forcibly targeting CaMKIIβ lacking the CTS to the centrosome by a different means should restore the ability of the CaMKIIβ mutant to inhibit dendrite growth and stimulate retraction. We therefore generated a chimeric protein in which we fused CaMKIIβ-RESΔCTS at its N terminus to the PACT (pericentrin-AKAP450 centrosomal targeting) domain, which localizes proteins to the centrosome (PACT-CaMKIIβ-RESΔCTS) (Fig. 3b and Supplementary Fig. 7b)[37]. Notably, addition of the PACT domain restored the ability of CaMKIIβ-RESΔCTS to restrict the elaboration of dendrites in the setting of CaMKIIβ RNAi (Fig. 3c). In complementary experiments, replacing the variable region of CaMKIIα with the CTS (CaMKIIα-CTS) conferred CaMKIIα with the ability to robustly restrict dendrite elaboration in the background of CaMKIIβ RNAi (Fig. 3b,f). Likewise, in contrast to wild-type CaMKIIα, expression of a PACT-CaMKIIα fusion protein restricted dendrite elaboration in the background of CaMKIIβ RNAi (Fig. 3b,f). Collectively, these results suggest that the CTS endows CaMKIIβ with the ability to localize to the centrosome and control dendrite patterning.

The finding that CaMKIIβ regulates dendrite patterning from the centrosome, where CaMKIIβ but not CaMKIIα is localized, suggested that centrosomal CaMKIIβ might function independently of CaMKIIα in the control of dendrite morphogenesis.

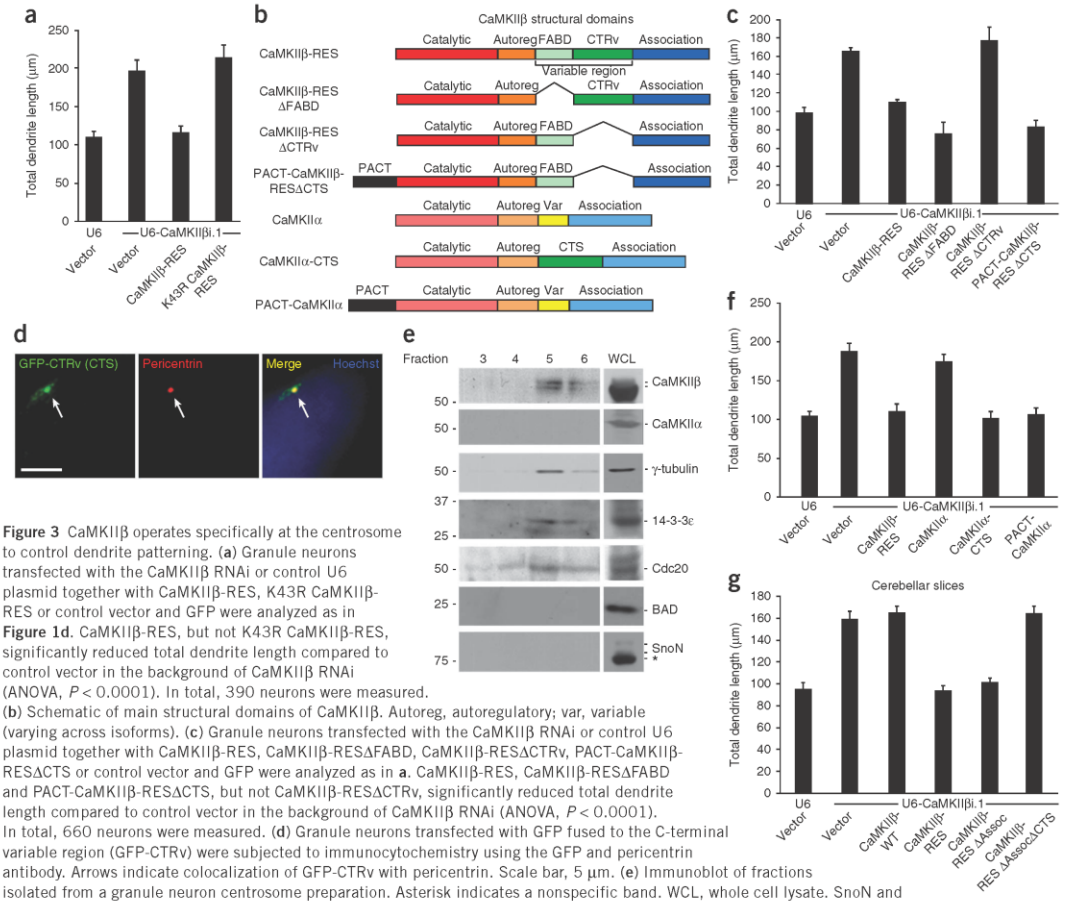


Figure 3 CaMKIIβ operates specifically at the centrosome to control dendrite patterning. (**a**) Granule neurons transfected with the CaMKIIβ RNAi or control U6 plasmid together with CaMKIIβ-RES, K43R CaMKIIβ-RES or control vector and GFP were analyzed as in **Figure 1d**. CaMKIIβ-RES, but not K43R CaMKIIβ-RES, significantly reduced total dendrite length compared to control vector in the background of CaMKIIβ RNAi (ANOVA, $P < 0.0001$). In total, 390 neurons were measured. (**b**) Schematic of main structural domains of CaMKIIβ. Autoreg, autoregulatory; var, variable (varying across isoforms). (**c**) Granule neurons transfected with the CaMKIIβ RNAi or control U6 plasmid together with CaMKIIβ-RES, CaMKIIβ-RESΔFABD, CaMKIIβ-RESΔCTRv, PACT-CaMKIIβ-RESΔCTS or control vector and GFP were analyzed as in **a**. CaMKIIβ-RES, CaMKIIβ-RESΔFABD and PACT-CaMKIIβ-RESΔCTS, but not CaMKIIβ-RESΔCTRv, significantly reduced total dendrite length compared to control vector in the background of CaMKIIβ RNAi (ANOVA, $P < 0.0001$). In total, 660 neurons were measured. (**d**) Granule neurons transfected with GFP fused to the C-terminal variable region (GFP-CTRv) were subjected to immunocytochemistry using the GFP and pericentrin antibody. Arrows indicate colocalization of GFP-CTRv with pericentrin. Scale bar, 5 μm. (**e**) Immunoblot of fractions isolated from a granule neuron centrosome preparation. Asterisk indicates a nonspecific band. WCL, whole cell lysate. SnoN and BAD mark nuclear and cytosolic fractions, respectively. (**f**) CaMKIIβ-RES, CaMKIIα-CTS and PACT-CaMKIIα, but not CaMKIIα, significantly reduced total dendrite length compared to control vector in the background of CaMKIIβ RNAi in granule neurons (ANOVA, $P < 0.0001$). In total, 600 neurons were measured. (**g**) CaMKIIβ-RES and CaMKIIβ-RESΔAssoc, but not CaMKIIβ-WT or CaMKIIβ-RESΔAssocΔCTS, significantly reduced granule neuron total dendrite length compared to control vector in the background of CaMKIIβ RNAi in cerebellar slices (ANOVA, $P < 0.0001$). In total, 479 neurons were measured. Error bars, s.e.m.

To test this hypothesis, we assessed the ability of a CaMKIIβ-RES mutant in which we removed the association domain (CaMKIIβ-RESΔAssoc) to restrict dendrite elaboration in the background of CaMKIIβ RNAi. In control biochemical analyses, deletion of the association domain abrogated the interaction of CaMKIIβ with CaMKIIα (**Supplementary Fig. 7c**). In morphology assays, deletion of the association domain had little or no effect on the ability of CaMKIIβ-RES to restrict the elaboration of dendrites in the background of CaMKIIβ RNAi in primary granule neurons and in cerebellar slices (**Fig. 3g** and **Supplementary Fig. 7d–f**), suggesting that the association domain is dispensable for CaMKIIβ control of dendrite patterning. Notably, as with full-length CaMKIIβ, deletion of the CTS or mutation of the ATP binding site (K43R) blocked the ability of the CaMKIIβΔAssoc mutant to restrict the elaboration of dendrites (**Fig. 3g**, **Supplementary Fig. 7d–f** and data not shown). In immunocytochemical analyses of granule neurons, although full-length GFP-CaMKIIβ was localized throughout the cell soma and processes (**Supplementary Fig. 7g**), CaMKIIβΔAssoc was enriched at the centrosome (**Supplementary Fig. 7h**). Deletion of the CTS blocked centrosomal enrichment of the CaMKIIβΔAssoc mutant (**Supplementary Fig. 7h**). Together, our data demonstrate that centrosomal CaMKIIβ functions independently of multimerization with CaMKIIα to regulate dendrite patterning.

The targeting protein PCM1 localizes CaMKIIβ to the centrosome

We next characterized the mechanism that localizes CaMKIIβ to the centrosome. In view of the importance of the CTS in the localization of CaMKIIβ at the centrosome, we reasoned that the CTS might associate with a protein that targets CaMKIIβ to the centrosome. We therefore performed a yeast two-hybrid assay using the CTS as bait and a human fetal brain library as prey to identify CTS-interacting proteins. In these assays, we identified the centrosomal targeting

protein PCM1 as an interactor of the CTS (data not shown). In coimmunoprecipitation analyses in cells, PCM1 formed a complex with CaMKIIβ but not CaMKIIα, and CaMKIIβ interacted with PCM1 in a CTS-dependent manner (**Fig. 4a**,**b**). In addition, endogenous PCM1 formed a complex with endogenous CaMKIIβ within centrosomal fractions of granule neurons (**Fig. 4c**).

To determine PCM1's role in targeting a pool of CaMKIIβ to the centrosome, we assessed the effect of PCM1 knockdown on the localization of CaMKIIβΔAssoc, which does not interact with CaMKIIα and is enriched at the centrosome (**Supplementary Fig. 7h**). We found that knockdown of endogenous PCM1, achieved efficiently by two shRNAs targeting distinct regions of PCM1, led to the loss of centrosomal enrichment of CaMKIIβΔAssoc in neurons (**Fig. 4d**,**e**). These results suggest that PCM1 specifically localizes CaMKIIβ to the centrosome.

The requirement for PCM1 in localizing CaMKIIβ to the centrosome led us next to determine the functional role of PCM1 in dendrite morphogenesis. We found that PCM1 knockdown substantially increased dendrite arborization and growth in primary granule neurons (**Fig. 4f**,**g** and data not shown). The PCM1

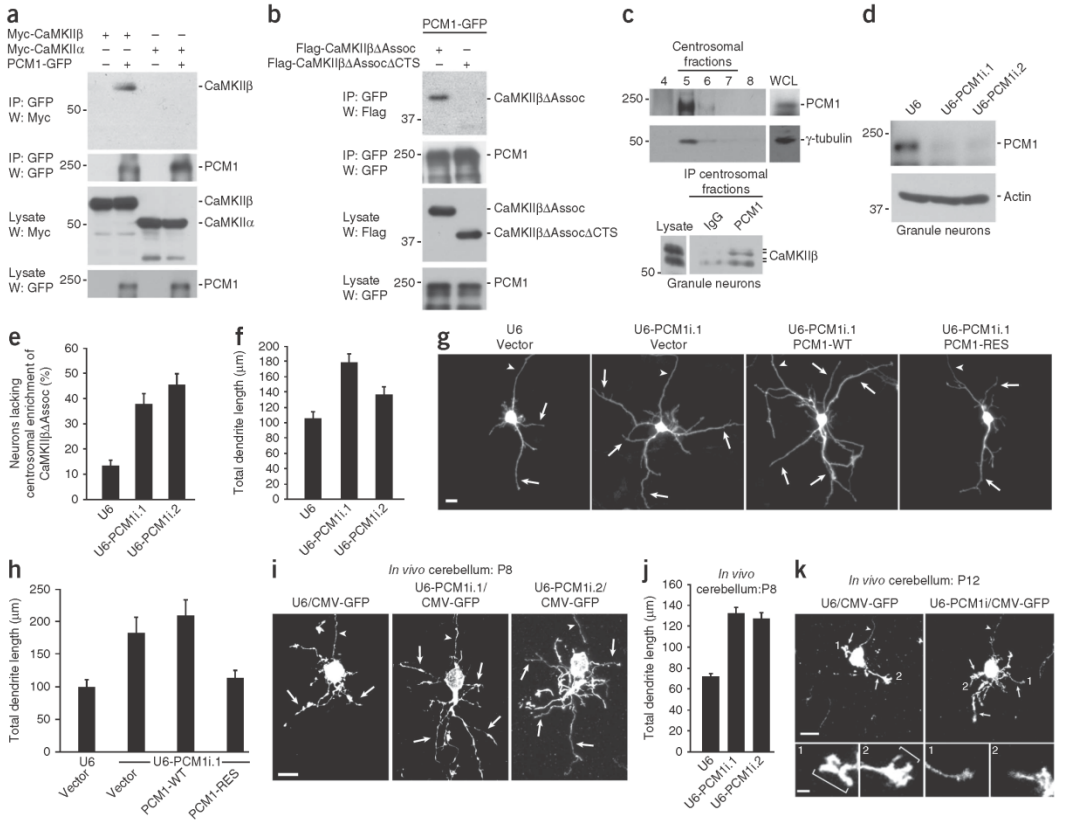


Figure 4 The centrosomal targeting protein PCM1 localizes CaMKIIβ to the centrosome. (**a**) Immunoblot (W) of lysates of 293T cells transfected with Myc-CaMKIIβ or Myc-CaMKIIα together with PCM1-GFP or control vector and immunoprecipitated (IP) using the GFP antibody. (**b**) Immunoblot of lysates of 293T cells transfected with Flag-CaMKIIβΔAssoc or Flag-CaMKIIβΔAssocΔCTS together with PCM1-GFP and immunoprecipitated using the GFP antibody. (**c**) Top panels, immunoblot of fractions isolated from a granule neuron centrosome preparation. Bottom panels, immunoblot of pooled centrosomal fractions immunoprecipitated with the PCM1 or control immunoglobulin G (IgG) antibody. (**d**) Immunoblot of lysates of granule neurons electroporated with the PCM1 RNAi or control U6 plasmid. (**e**) Granule neurons transfected with a PCM1 RNAi or control U6 plasmid together with GFP-CaMKIIβΔAssoc were analyzed as in **Figure 3d**. The percentage of neurons lacking centrosomal enrichment of GFP-CaMKIIβΔAssoc was significantly greater in PCM1 knockdown neurons than in control U6-transfected neurons (ANOVA, $P < 0.005$). In total, 394 neurons were analyzed. (**f**) Total dendrite length was significantly greater in PCM1 knockdown neurons than in control U6-transfected neurons (ANOVA, $P < 0.0001$). In total, 180 neurons were measured. (**g**) Granule neurons transfected with the PCM1 RNAi or control U6 plasmid together with PCM1-WT, PCM1-RES or control vector and GFP were analyzed as in **f**. Scale bar, 10 μm. (**h**) PCM1-RES, but not PCM1-WT, significantly reduced total dendrite length compared to control vector in the background of PCM1 RNAi (ANOVA, $P < 0.0001$). In total, 240 neurons were measured. (**i**) Rat pups electroporated *in vivo* with a U6-PCM1i/CMV-GFP RNAi or control U6/CMV-GFP plasmid were killed at P8 and analyzed as in **Figure 2d**. Scale bar, 10 μm. (**j**) Total dendrite length was significantly greater in IGL granule neurons in PCM1 knockdown rats than in control U6 rats (ANOVA, $P < 0.0001$). In total, 270 neurons were measured. (**k**) Rat pups electroporated *in vivo* with the U6-PCM1i–CMV-GFP RNAi or control U6–CMV-GFP plasmid were killed at P12 and analyzed as in **Figure 2g**. Scale bar, 10 μm. Inset: scale bar, 2.5 μm. Error bars, s.e.m.

RNAi-induced dendrite phenotype was reversed by expression of PCM1 encoded by an RNAi-resistant cDNA (PCM1-RES) but not wild-type cDNA (PCM1-WT) (**Fig. 4g**,**h** and data not shown), indicating the specificity of the PCM1 RNAi-induced phenotype. These results suggest that, like CaMKIIβ, PCM1 induces dendrite retraction. We also determined the effect of PCM1 knockdown on dendrite morphogenesis in the cerebellar cortex *in vivo*. We found that PCM1 knockdown triggered robust elaboration of IGL granule neuron dendrite arbors in P8 rat pups (**Fig. 4i**,**j** and **Supplementary Fig. 8a**). Analyses in P12 rat pups revealed that granule neuron dendrite arbors in PCM1 knockdown rats remained long and highly branched, containing substantially fewer dendritic claws than those in control rats (**Fig. 4k** and **Supplementary Fig. 8b–d**). Thus, PCM1 knockdown phenocopied the effect of CaMKIIβ knockdown in impairing dendrite pruning and blocking dendrite differentiation at a stage of exuberant arbors *in vivo*. In other experiments, PCM1 knockdown suppressed the ability of activated CaMKIIβ to stimulate dendrite retraction in granule neurons (**Supplementary Fig. 8e**). Collectively, these results suggest that PCM1 is required for CaMKIIβ localization to the centrosome and hence the ability of CaMKIIβ to regulate the patterning of dendrites.

The ubiquitin ligase Cdc20-APC is a novel CaMKIIβ substrate

As the catalytic activity of CaMKIIβ at the centrosome was required for regulating dendrite morphogenesis, we next sought to identify the protein substrates of centrosomal CaMKIIβ in neurons. We considered proteins that have established roles in dendrite development, localize to the centrosome, and contain a consensus sequence for CaMKII phosphorylation. Recently, the major mitotic E3 ubiquitin ligase Cdc20-APC has been identified as a critical mediator of dendrite growth that operates at the centrosome in postmitotic neurons[38]. We found that the N terminus of the APC coactivator Cdc20 had several potential CaMKII sites, including Ser51 and Ser86, which conform to the CaMKII consensus sequence[39]. Consistent with the possibility that CaMKIIβ might phosphorylate Cdc20, CaMKIIβ and Cdc20 formed a complex in 293T cells and neurons (**Fig. 5a**,**b** and **Supplementary Fig. 9a**).

Purified CaMKII catalyzed the phosphorylation of full-length and an N-terminal domain of Cdc20 (1–101) *in vitro*, and mass spectrometric analyses of phosphorylated Cdc20 revealed that CaMKII phosphorylated Cdc20 at Ser51, Ser84 and Ser86 (data not shown). Mutational analysis demonstrated that Ser86 and Ser51 contributed to CaMKII-mediated phosphorylation of Cdc20, whereas mutation

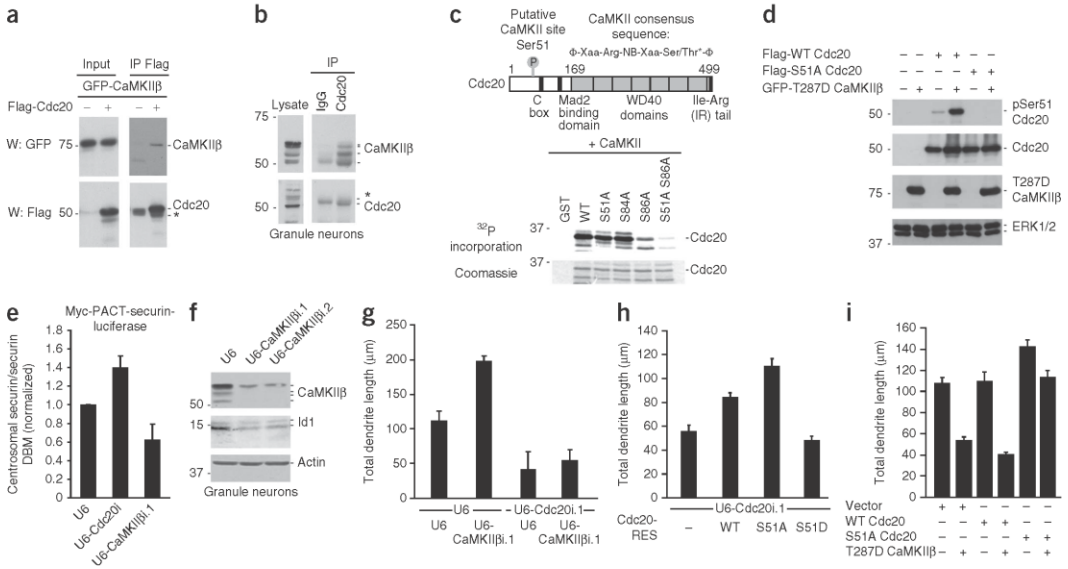


Figure 5 CaMKIIβ phosphorylates Cdc20 at Ser51 and thereby inhibits centrosomal Cdc20-APC activity in neurons. (**a**) Immunoblot (W) of lysates of 293T cells transfected with GFP-CaMKIIβ together with Flag-Cdc20 or control vector and immunoprecipitated (IP) using the Flag antibody. Asterisk, IgG heavy chain. (**b**) Immunoblot of lysates of granule neurons immunoprecipitated with the Cdc20 or control (IgG) antibody. Asterisk, IgG heavy chain. (**c**) Top, schematic of CaMKII consensus sequence. Asterisk, the phosphorylated serine or threonine; Φ, hydrophobic amino acid; NB, non-basic amino acid. Bottom, phosphorylation of recombinant wild-type (WT) and S51A, S84A, S86A and S51A S86A mutants of an N-terminal region of Cdc20 (1–101) fused to glutathione S-transferase (GST) upon incubation *in vitro* with purified CaMKII and ^{32}P-ATP. (**d**) Immunoblot of lysates of 293T cells transfected with Flag-WT Cdc20 or Flag-S51A Cdc20 together with GFP-T287D CaMKIIβ or control vector. (**e**) Granule neurons were transfected with the Cdc20 RNAi, CaMKIIβ RNAi or control U6 plasmid together with Myc-PACT-securin-luciferase or Myc-PACT-securin-DBM-luciferase and analyzed by luminometry. The relative amount of the centrosomal securin-luciferase reporter was significantly greater in Cdc20 knockdown neurons and significantly lower in CaMKIIβ knockdown neurons than in control U6-transfected neurons (Kruskal–Wallis test, $P < 0.01$, $n = 4$). (**f**) Immunoblot of lysates of granule neurons electroporated with the CaMKIIβ RNAi or control U6 plasmid. (**g**) CaMKIIβ knockdown significantly increased total dendrite length compared to control, and Cdc20 RNAi significantly reduced total dendrite length in granule neurons in the presence or absence of CaMKIIβ RNAi (ANOVA, $P < 0.0001$). In total, 420 neurons were measured. (**h**) Expression of S51A Cdc20-RES, but not S51D Cdc20-RES, significantly increased total dendrite length compared to control vector or WT Cdc20-RES in the background of Cdc20 RNAi in granule neurons (ANOVA, $P < 0.0001$). In total, 428 neurons were measured. (**i**) Expression of T287D CaMKIIβ significantly reduced total dendrite length compared to control in the background of control vector or WT Cdc20, but not in the background of S51A Cdc20 (ANOVA, $P < 0.0001$). In total, 540 neurons were measured. Error bars, s.e.m.

of Ser84 to alanine did not appreciably alter ^{32}P-ATP incorporation (Fig. 5c), suggesting that CaMKIIβ does not stoichiometrically phosphorylate Ser84 *in vitro*. Notably, Ser51 but not Ser86 is evolutionarily conserved from zebrafish to humans. We therefore generated an antibody to specifically recognize Cdc20 phosphorylated at Ser51. The phospho-Cdc20 antibody recognized wild-type Cdc20 but not the Ser51 mutant of Cdc20 when Cdc20 was expressed with activated CaMKIIβ in cells (Fig. 5d). Immunoblotting experiments using the phospho-Cdc20 antibody revealed that endogenous Cdc20 was phosphorylated at Ser51 in granule neurons, and knockdown of CaMKIIβ reduced endogenous Ser51-phosphorylated Cdc20 in neurons (Supplementary Fig. 9b). Together, these results indicate that endogenous CaMKIIβ phosphorylates Cdc20 at Ser51 in neurons, raising the question of whether CaMKIIβ regulates the activity of Cdc20-APC at the centrosome to control dendrite patterning.

To determine the role of CaMKIIβ in regulating the ubiquitin ligase activity of centrosomal Cdc20-APC in neurons, we first used a Myc-PACT-securin-luciferase reporter, which is a validated readout of centrosomal Cdc20-APC activity in neurons[38]. The reporter encodes the first 88 amino acids of the mitotic Cdc20-APC substrate securin fused to firefly luciferase. This portion of securin harbors the destruction box (D-box), an APC degron[40], and the PACT domain localizes the reporter to the centrosome[38]. A mutant version of the reporter containing a mutation in the securin D-box (securin-DBM) serves as control[40]. Thus, the ratio of securin to securin-DBM luciferase is inversely proportional to centrosomal Cdc20-APC activity[38]. As reported, Cdc20 knockdown increased the relative levels of the securin-luciferase reporter in granule neurons (Fig. 5e)[38]. By contrast, we found that CaMKIIβ knockdown reduced relative securin-luciferase reporter in neurons (Fig. 5e), suggesting that CaMKIIβ inhibits centrosomal Cdc20-APC activity in neurons. CaMKIIβ knockdown also reduced levels of the centrosomal helix-loop-helix protein Id1 (inhibitor of DNA 1) (Fig. 5f), which is a physiological substrate of centrosomal Cdc20-APC in neurons[38]. Collectively, these data suggest that CaMKIIβ inhibits Cdc20-APC activity at the centrosome and the consequent degradation of its substrates in neurons.

In view of Cdc20-APC's function in dendrite growth and arborization, the finding that CaMKIIβ inhibits the activity of Cdc20-APC at the centrosome in neurons suggested that CaMKIIβ might stimulate dendrite retraction by regulating centrosomal Cdc20-APC activity. Consistent with this prediction, in epistasis analyses, Cdc20 knockdown suppressed the effect of CaMKIIβ knockdown on dendrite elaboration in granule neurons (Fig. 5g). In addition, knockdown of the centrosomal Cdc20-APC substrate Id1 suppressed the ability of CaMKIIβ to induce dendrite retraction (data not shown). Together, these results suggest that CaMKIIβ stimulates dendrite retraction by inhibiting centrosomal Cdc20-APC activity in neurons.

Phosphorylation induces Cdc20 dispersion from the centrosome

We next assessed the role of Cdc20 phosphorylation at Ser51 in CaMKIIβ-regulated dendrite patterning. Expression of Cdc20-RES promotes dendrite growth and elaboration in the background of Cdc20 RNAi[38]. We found that expression of a mutant Cdc20-RES protein in which Ser51 was replaced with the non-phosphorylatable residue alanine (S51A Cdc20-RES) stimulated dendrite growth more effectively than Cdc20-RES (Fig. 5h). By contrast, expression of a mutant Cdc20-RES protein in which Ser51 was replaced with phosphomimetic residue aspartate (S51D Cdc20-RES) failed to stimulate dendrite growth (Fig. 5h). Notably, expression of S51A Cdc20 but not wild-type Cdc20 substantially impaired the ability of constitutively active CaMKIIβ to induce dendrite retraction (Fig. 5i). In other experiments, mutation of Ser84 or Ser86 had little or no effect on the ability of Cdc20-RES to drive dendrite elaboration (Supplementary Fig. 9c). Collectively, our results suggest that CaMKIIβ phosphorylates Cdc20 at Ser51 and inhibits centrosomal Cdc20-APC activity in neurons, thereby triggering a switch from dendrite growth and arborization to dendrite retraction.

We next determined the mechanism by which CaMKIIβ-induced phosphorylation of Cdc20 inhibits Cdc20-APC function at the centrosome in neurons. Because the localization of Cdc20 at the centrosome is critical for its function in dendrite growth and arborization[38], we asked whether CaMKIIβ phosphorylation of Cdc20 might interfere with the centrosomal localization of Cdc20. Indeed, expression of wild-type or constitutively active CaMKIIβ (T287D), but not the catalytically inactive K43R CaMKIIβ mutant or CaMKIIα, induced dispersion of endogenous Cdc20 from the centrosome in granule neurons (Fig. 6a,b and Supplementary Fig. 9d,e). Expression of CaMKIIβ had little or no effect on centrosomal structure in neurons as monitored by endogenous pericentrin or co-transfection with GFP-centrin (Supplementary Fig. 9e and data not shown). Conversely, CaMKIIβ knockdown reduced basal Cdc20 dispersion in granule neurons (Supplementary Fig. 9f), suggesting that endogenous CaMKIIβ inhibits the centrosomal localization of Cdc20 in neurons.

To assess whether CaMKIIβ stimulates Cdc20 dispersion independently of CaMKIIα in neurons, we tested the effect of deletion of the association domain on the ability of CaMKIIβ to induce endogenous Cdc20 dispersion in granule neurons. The CaMKIIβΔAssoc mutant protein induced Cdc20 dispersion as effectively as CaMKIIβ (Fig. 6c), suggesting that the association domain is dispensable for CaMKIIβ-induced dispersion of Cdc20 in neurons. Deletion of the CTS in CaMKIIβΔAssoc diminished its ability to induce Cdc20 dispersion (Fig. 6c). Together, our results suggest that the protein kinase CaMKIIβ functions at the centrosome independently of CaMKIIα to drive Cdc20 dispersion in neurons.

To determine whether the regulation of Cdc20 localization by CaMKIIβ is relevant to dendrite morphogenesis, we first compared dendrite length in granule neurons showing centrosomal enrichment of Cdc20 with neurons in which Cdc20 was dispersed. Neurons with dispersed endogenous Cdc20 had substantially shorter dendrites than neurons with centrosomally localized Cdc20 (35.8 ± 10.2 μm in neurons with dispersed Cdc20 versus 105.5 ± 9.9 μm in neurons with centrosomal Cdc20; *t*-test, $P < 0.01$) (Fig. 6d and Supplementary Fig. 9g). In other experiments, dispersion of endogenous Cdc20 in granule neurons increased with maturation (Supplementary Fig. 9h), suggesting that Cdc20 dispersion correlates temporally with CaMKIIβ function in dendrite retraction and pruning. If Cdc20 dispersion from the centrosome is crucial for CaMKIIβ-induced dendrite retraction, then forcibly localizing Cdc20 to the centrosome should override the CaMKIIβ response. In control experiments, we found that although expression of T287D CaMKIIβ dispersed Cdc20, T287D CaMKIIβ failed to disperse a Cdc20 protein fused to the centrosome-localizing PACT domain (PACT-Cdc20) (Fig. 6e), suggesting that PACT-Cdc20 is insensitive to CaMKIIβ regulation. Notably, in morphology assays, expression of PACT-Cdc20, but not wild-type Cdc20, suppressed the ability of T287D CaMKIIβ to promote dendrite retraction in granule neurons (Fig. 6f). Collectively, these results suggest that Cdc20 dispersion from the centrosome is critical for CaMKIIβ regulation of dendrite patterning.

We next assessed the role of CaMKIIβ-induced Cdc20 phosphorylation at Ser51 in the control of Cdc20 localization to the centrosome. In immunocytochemical analyses using the phospho-Cdc20

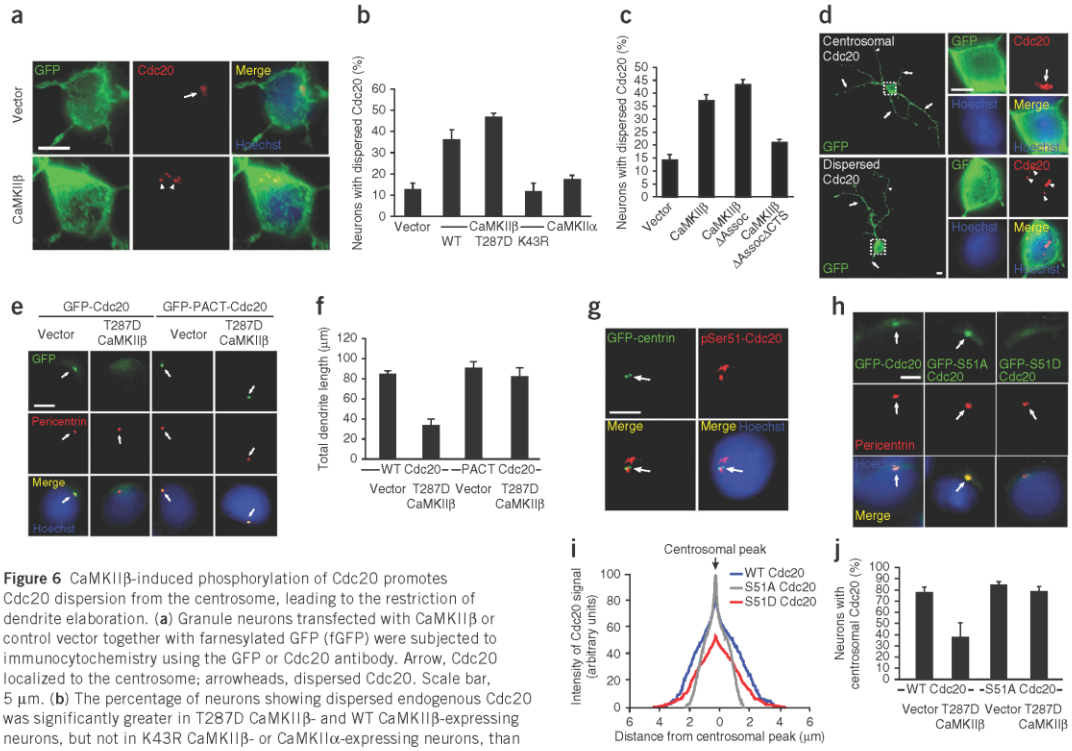


Figure 6 CaMKIIβ-induced phosphorylation of Cdc20 promotes Cdc20 dispersion from the centrosome, leading to the restriction of dendrite elaboration. (**a**) Granule neurons transfected with CaMKIIβ or control vector together with farnesylated GFP (fGFP) were subjected to immunocytochemistry using the GFP or Cdc20 antibody. Arrow, Cdc20 localized to the centrosome; arrowheads, dispersed Cdc20. Scale bar, 5 μm. (**b**) The percentage of neurons showing dispersed endogenous Cdc20 was significantly greater in T287D CaMKIIβ- and WT CaMKIIβ-expressing neurons, but not in K43R CaMKIIβ- or CaMKIIα-expressing neurons, than in control vector-transfected neurons (ANOVA, $P < 0.0001$). In total, 622 neurons were analyzed. (**c**) The percentage of neurons showing dispersed endogenous Cdc20 was significantly greater in CaMKIIβ- and CaMKIIβΔAssoc-expressing neurons, but not in CaMKIIβΔAssocΔCTS-expressing neurons, than in control vector–transfected neurons (ANOVA, $P < 0.0001$). In total, 360 neurons were analyzed. (**d**) Granule neurons transfected with T287D CaMKIIβ together with fGFP were analyzed as in **a**. Arrows indicate dendrites, whereas the dashed boxes indicate the magnified regions in inset images. Scale bar, 5 μm. (**e**) Granule neurons transfected with GFP-Cdc20 or GFP-PACT-Cdc20 together with T287D CaMKIIβ or control vector were analyzed as in **Figure 3d**. Arrows indicate the location of the centrosome, which is labeled with pericentrin. Scale bar, 5 μm. (**f**) Total dendrite length was significantly lower in T287D CaMKIIβ-expressing neurons than in control vector–transfected neurons in the background of Cdc20 (ANOVA, $P < 0.0005$), but not in the background of PACT-Cdc20. In total, 360 neurons were measured. (**g**) Granule neurons transfected with GFP-centrin were subjected to immunocytochemistry using the GFP or phosphoSer51-Cdc20 antibody. Arrows, centrosome (labeled with GFP-centrin). Scale bar, 5 μm. (**h**) Granule neurons transfected with GFP-Cdc20, GFP-S51A Cdc20 or GFP-S51D Cdc20 were analyzed as in **e**. Arrows, centrosome (labeled with pericentrin). Scale bar, 5 μm. (**i**) Line scan analysis of granule neurons treated as in **h**. Centrosome location was identified using pericentrin immunoreactivity. S51A Cdc20 had enhanced signal at the centrosome and reduced cytosolic signal compared to WT Cdc20. By contrast, S51D Cdc20 had reduced centrosomal signal and broader non-centrosomal signal compared to WT Cdc20 (based on peak centrosomal signal intensity, ANOVA, $P < 0.0001$). In total, 270 neurons were analyzed. (**j**) Expression of T287D CaMKIIβ significantly reduced the percentage of neurons with centrosomally enriched WT Cdc20 compared to control vector, but had little or no effect on the centrosomal localization of S51A Cdc20 (ANOVA, $P < 0.01$). In total, 361 neurons were analyzed. Error bars, s.e.m.

antibody, we found that endogenous Ser51-phosphorylated Cdc20 immunoreactivity had a broader distribution beyond the centrosome in granule neurons, suggesting that the dispersed pool of Cdc20 is phosphorylated at Ser51 (**Fig. 6g**). Expression of activated CaMKIIβ increased the number of neurons harboring phosphoSer51-Cdc20 immunofluorescence (**Supplementary Fig. 9i**), whereas knockdown of endogenous CaMKIIβ but not CaMKIIα reduced the number of neurons with Ser51-phosphorylated Cdc20 (**Supplementary Fig. 9j**). To directly test the importance of Ser51 phosphorylation in Cdc20 dispersion, we characterized the effect of mutations of Ser51 on Cdc20 localization in neurons. We found that the non-phosphorylatable S51A Cdc20 mutant protein seemed to be more enriched at the centrosome than wild-type Cdc20, whereas the phosphomimetic S51D mutant seemed to be less enriched at the centrosome (**Fig. 6h**). Quantification of the Cdc20 immunofluorescence signal using line scan analyses revealed that the S51A Cdc20 signal peaked at the centrosome and showed a more limited distribution outside the centrosome (**Fig. 6i**). By contrast, S51D Cdc20 had a widened centrosomal peak with a broader non-centrosomal signal (**Fig. 6h,i**). In other experiments, expression of T287D CaMKIIβ robustly induced dispersion of wild-type Cdc20 but not S51A Cdc20 from the centrosome in granule neurons (**Fig. 6j**), suggesting CaMKIIβ induces Cdc20 dispersion in a Ser51 phosphorylation-dependent manner. Notably, addition of the PACT domain to S51D Cdc20 restored the ability of S51D Cdc20 to stimulate dendrite growth and elaboration, suggesting that the function of Ser51 phosphorylation

is to localize Cdc20 away from the centrosome (Supplementary Fig. 9k). These data suggest that CaMKIIβ-induced phosphorylation of Cdc20 at Ser51 triggers its dispersion from the centrosome, leading to the inhibition of centrosomal Cdc20-APC activity and the promotion of dendrite retraction in neurons. Collectively, we have elucidated a centrosomal CaMKIIβ signaling pathway that controls the patterning of dendrite arbors with important consequences for the establishment of neuronal connectivity in the mammalian brain (see model in Supplementary Fig. 10).

DISCUSSION

In this study, we have discovered the first isoform-specific catalytic function of CaMKIIβ in the mammalian brain. CaMKIIβ operates at the centrosome in a CaMKIIα-independent manner to drive dendrite retraction and pruning. CaMKIIβ localizes to the centrosome by forming a complex with the centrosomal trafficking protein PCM1. Accordingly, PCM1 is required for CaMKIIβ regulation of dendrite morphogenesis and the pruning of dendrites in postnatal rat pups *in vivo*. We have also identified the ubiquitin ligase Cdc20-APC, which is enriched at the centrosome in neurons, as a new substrate of CaMKIIβ. CaMKIIβ phosphorylates Cdc20 at Ser51 and thereby triggers the dispersion of Cdc20 from the centrosome, culminating in the inhibition of centrosomal Cdc20-APC activity and consequent dendrite retraction. Collectively, our findings define a novel isoform-specific function of CaMKIIβ that regulates ubiquitin-dependent protein degradation at the centrosome and thereby orchestrates dendrite patterning in mammalian brain neurons.

The identification of an unexpected function for CaMKIIβ in dendrite retraction and pruning has considerable ramifications for our understanding of the major protein kinase CaMKII. As the predominant CaMKII isoforms in the brain, CaMKIIα and CaMKIIβ form holoenzyme complexes in neurons[16–18]. However, nearly all of the reported functions of CaMKII have focused on CaMKIIα in homomeric or heteromeric complexes[19–22]. Accordingly, CaMKIIβ has been previously relegated to a largely redundant role within CaMKIIα heteromeric complexes or as a scaffold recruiting CaMKIIα to specific cellular locales such as dendritic spines[34–36]. Our finding that CaMKIIβ operates at the centrosome in a CaMKIIα-independent manner unveils a biological function for CaMKIIβ homomeric complexes. Thus, rather than simply contributing to CaMKIIα complexes, CaMKIIβ homomers have a major biological function at the centrosome as drivers of dendrite retraction and pruning and consequent neuronal connectivity in the mammalian brain. During brain development, CaMKIIβ expression peaks earlier than CaMKIIα[41]. Therefore, it will be interesting to determine in future studies whether centrosomal CaMKIIβ homomeric complexes might also contribute to earlier aspects of neuronal development in which centrosomal signaling is thought to be critical, including neuronal polarization and migration[42,43]. Beyond the nervous system, it will be important to identify the role of CaMKIIβ homomers in cellular homeostasis and development in other organ systems.

Elucidation of the CaMKIIβ–Cdc20 signaling link at the centrosome provides a mechanism by which CaMKIIβ promotes dendrite retraction and pruning. By inducing the phosphorylation of Cdc20 and inhibiting Cdc20-APC ubiquitin ligase activity at the centrosome, CaMKIIβ triggers a switch from dendrite growth and elaboration to dendrite retraction and pruning. These findings suggest that the centrosome may represent a critical signaling platform that integrates diverse cellular signals to determine the phase of dendrite morphogenesis in neurons. In view of the fundamental role of centrosome signaling in diverse systems from the control of cell polarity and migration to ciliary morphogenesis to vesicular transport[43], it will be interesting to explore whether centrosomal CaMKIIβ phosphorylation of Cdc20, or of other substrates that have yet to be identified, might contribute to these diverse biological processes.

The identification of the CaMKIIβ–Cdc20 signaling link also advances our understanding of the regulation of the major ubiquitin ligase Cdc20-APC. CaMKIIβ phosphorylates Cdc20 at Ser51, inducing dispersion of Cdc20 from the centrosome and inhibiting centrosomal Cdc20-APC activity. The phosphorylation-dependent control of the subcellular localization of Cdc20 represents a new mode of regulation of the ubiquitin ligase Cdc20-APC. Of note, CaMKIIβ had little or no effect on Cdc20 binding to core APC subunits, the *in vitro* ubiquitin ligase activity of Cdc20-APC, or Cdc20 polyubiquitylation (data not shown), highlighting the significance of Cdc20 dispersion as a crucial cellular mechanism for regulating centrosomal Cdc20-APC activity. Future research should explore whether Cdc20 has more roles outside of the centrosome, bearing in mind the possibility that Cdc20 dispersion might be part of a developmental program that coordinates dendrite patterning with other steps in the establishment of neuronal connectivity.

Although CaMKIIβ and CaMKIIα are similar in their catalytic, autoregulatory and association domains, they diverge in their variable region[41]. We have identified a CTS in the unique variable region of CaMKIIβ that provides spatial specificity for CaMKIIβ function. Further, we have found that the CTS mediates the interaction of CaMKIIβ specifically with the centrosomal targeting protein PCM1. Although PCM1 targets structural proteins to the centrosome[44], our findings suggest that PCM1 may also drive the centrosomal localization of signaling proteins. Notably, in addition to interacting with CaMKIIβ, endogenous PCM1 formed a complex with endogenous Cdc20 in neurons (data not shown). These observations suggest that beyond recruiting CaMKIIβ to the centrosome, PCM1 may also operate as a scaffold protein that organizes the CaMKIIβ–Cdc20 signaling pathway at the centrosome.

Although signaling and cell-intrinsic mechanisms that promote dendrite growth have been the subject of substantial interest[5–8], the master regulatory mechanisms that govern the developmental transition from dendrite elaboration to dendrite pruning in the mammalian brain have remained to be elucidated. The identification of centrosomal CaMKIIβ signaling as a mechanism that restricts dendrite elaboration and promotes dendrite pruning suggests that pathways that actively drive dendrite retraction have evolved to sculpt dendrite arbors and thus establish accurate neuronal circuits. As abnormalities in dendrite development represent prominent pathological features in mental retardation and autism spectrum disorders[4,45,46], it will be interesting to determine whether deregulation of centrosomal CaMKIIβ signaling contributes to the pathogenesis of neurodevelopmental disorders of cognition.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/natureneuroscience/.

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

We thank M. Ericsson for assistance with electron microscopy experiments and J. Blenis, G. Corfas and members of the Bonni laboratory for discussions and critical reading of the manuscript. This work was supported by US National Institutes of Health grant NS051255 (A.B.), a Ruth L. Kirschstein National Research Service Award (National Cancer Institute), a Brain Science Foundation grant (A.H.K.) and a Human Frontier Science Program Long-term Fellowship (Y.I.).

AUTHOR CONTRIBUTIONS
A.B. directed and coordinated the project. S.V.P., A.H.K. and Y.I. designed and performed *in vivo* experiments, biochemical assays and morphological analyses. J.T.W.-G. prepared mass spectrometry samples and completed analyses of Cdc20 phosphorylation in S.P.G.'s laboratory. A.M. contributed molecular reagents. The manuscript was written by S.V.P. and A.B. and critically reviewed and commented on by all authors.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Published online at http://www.nature.com/natureneuroscience/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Altman, J. & Bayer, S. *Development of the Cerebellar System: In Relation to its Evolution, Structure, and Functions* (CRC, New York, 1997).
2. Palay, S. & Chan-Palay, V. *Cerebellar Cortex: Cytology and Organization* (Springer, New York, 1974).
3. Lee, A. *et al.* Control of dendritic development by the *Drosophila* fragile X-related gene involves the small GTPase Rac1. *Development* **130**, 5543–5552 (2003).
4. Kaufmann, W.E. & Moser, H.W. Dendritic anomalies in disorders associated with mental retardation. *Cereb. Cortex* **10**, 981–991 (2000).
5. Corty, M.M., Matthews, B.J. & Grueber, W.B. Molecules and mechanisms of dendrite development in *Drosophila*. *Development* **136**, 1049–1061 (2009).
6. Miller, F.D. & Kaplan, D.R. Signaling mechanisms underlying dendrite formation. *Curr. Opin. Neurobiol.* **13**, 391–398 (2003).
7. Jan, Y.N. & Jan, L.Y. The control of dendrite development. *Neuron* **40**, 229–242 (2003).
8. Konur, S. & Ghosh, A. Calcium signaling and the control of dendritic development. *Neuron* **46**, 401–405 (2005).
9. Hudmon, A. & Schulman, H. Structure-function of the multifunctional Ca^{2+}/calmodulin-dependent protein kinase II. *Biochem. J.* **364**, 593–611 (2002).
10. Wayman, G.A., Lee, Y.-S., Tokumitsu, H., Silva, A. & Soderling, T.R. Calmodulin-kinases: modulators of neuronal development and plasticity. *Neuron* **59**, 914–931 (2008).
11. Erondu, N.E. & Kennedy, M.B. Regional distribution of type II Ca^{2+}/calmodulin-dependent protein kinase in rat brain. *J. Neurosci.* **5**, 3270–3277 (1985).
12. Hoelz, A., Nairn, A.C. & Kuriyan, J. Crystal structure of a tetradecameric assembly of the association domain of Ca^{2+}/calmodulin-dependent kinase II. *Mol. Cell* **11**, 1241–1251 (2003).
13. Rosenberg, O.S. *et al.* Oligomerization states of the association domain and the holoenyzme of Ca^{2+}/CaM kinase II. *FEBS J.* **273**, 682–694 (2006).
14. Kolb, S.J., Hudmon, A., Ginsberg, T.R. & Waxham, M.N. Identification of domains essential for the assembly of calcium/calmodulin-dependent protein kinase II holoenzymes. *J. Biol. Chem.* **273**, 31555–31564 (1998).
15. Griffith, L.C. Calcium/calmodulin-dependent protein kinase II: an unforgettable kinase. *J. Neurosci.* **24**, 8391–8393 (2004).
16. Kanaseki, T., Ikeuchi, Y., Sugiura, H. & Yamauchi, T. Structural features of Ca^{2+}/calmodulin-dependent protein kinase II revealed by electron microscopy. *J. Cell Biol.* **115**, 1049–1060 (1991).
17. Vallano, M.L. Separation of isozymic forms of type II calcium/calmodulin-dependent protein kinase using cation-exchange chromatography. *J. Neurosci. Methods* **30**, 1–9 (1989).
18. Miller, S.G. & Kennedy, M.B. Distinct forebrain and cerebellar isozymes of type II Ca^{2+}/calmodulin-dependent protein kinase associate differently with the postsynaptic density fraction. *J. Biol. Chem.* **260**, 9039–9046 (1985).
19. Colbran, R.J. Targeting of calcium/calmodulin-dependent protein kinase II. *Biochem. J.* **378**, 1–16 (2004).
20. Nicoll, R.A. & Malenka, R.C. Contrasting properties of two forms of long-term potentiation in the hippocampus. *Nature* **377**, 115–118 (1995).
21. Zou, D.J. & Cline, H.T. Postsynaptic calcium/calmodulin-dependent protein kinase II is required to limit elaboration of presynaptic and postsynaptic neuronal arbors. *J. Neurosci.* **19**, 8909–8918 (1999).
22. Silva, A.J., Paylor, R., Wehner, J.M. & Tonegawa, S. Impaired spatial learning in alpha-calcium-calmodulin kinase II mutant mice. *Science* **257**, 206–211 (1992).
23. Powell, S.K., Rivas, R.J., Rodriguez-Boulan, E. & Hatten, M.E. Development of polarity in cerebellar granule neurons. *J. Neurobiol.* **32**, 223–236 (1997).
24. Hatten, M.E. & Heintz, N. Mechanisms of neural patterning and specification in the developing cerebellum. *Annu. Rev. Neurosci.* **18**, 385–408 (1995).
25. Mason, C.A., Morrison, M.E., Ward, M.S., Zhang, Q. & Baird, D.H. Axon-target interactions in the developing cerebellum. *Perspect. Dev. Neurobiol.* **5**, 69–82 (1997).
26. Gaudillière, B., Shi, Y. & Bonni, A. RNA interference reveals a requirement for myocyte enhancer factor 2A in activity-dependent neuronal survival. *J. Biol. Chem.* **277**, 46442–46446 (2002).
27. Gaudillière, B., Konishi, Y., de la Iglesia, N., Yao, G. & Bonni, A.A. CaMKII-NeuroD signaling pathway specifies dendritic morphogenesis. *Neuron* **41**, 229–241 (2004).
28. Konishi, Y., Stegmuller, J., Matsuda, T., Bonni, S. & Bonni, A. Cdh1-APC controls axonal growth and patterning in the mammalian brain. *Science* **303**, 1026–1030 (2004).
29. Ramon y Cajal, S. The cerebellum. in *Histology of the Nervous System Vol. II* (eds. Swanson, N. & Swanson, L.) 3–124 (Oxford Univ. Press, New York, 1995).
30. de la Torre-Ubieta, L. *et al.* A FOXO-Pak1 transcriptional pathway controls neuronal polarity. *Genes Dev.* **24**, 799–813 (2010).
31. Shalizi, A. *et al.* PIASx is a MEF2 SUMO E3 ligase that promotes postsynaptic dendritic morphogenesis. *J. Neurosci.* **27**, 10037–10046 (2007).
32. Shalizi, A. *et al.* A calcium-regulated MEF2 sumoylation switch controls postsynaptic differentiation. *Science* **311**, 1012–1017 (2006).
33. Vaillant, A.R. *et al.* Signaling mechanisms underlying reversible, activity-dependent dendrite formation. *Neuron* **34**, 985–998 (2002).
34. Okamoto, K., Narayanan, R., Lee, S.H., Murata, K. & Hayashi, Y. The role of CaMKII as an F-actin-bundling protein crucial for maintenance of dendritic spine structure. *Proc. Natl. Acad. Sci. USA* **104**, 6418–6423 (2007).
35. Fink, C.C. *et al.* Selective regulation of neurite extension and synapse formation by the beta but not the alpha isoform of CaMKII. *Neuron* **39**, 283–297 (2003).
36. O'Leary, H., Lasda, E. & Bayer, K.U. CaMKIIβ association with the actin cytoskeleton is regulated by alternative splicing. *Mol. Biol. Cell* **17**, 4656–4665 (2006).
37. Gillingham, A.K. & Munro, S. The PACT domain, a conserved centrosomal targeting motif in the coiled-coil proteins AKAP450 and pericentrin. *EMBO Rep.* **1**, 524–529 (2000).
38. Kim, A.H. *et al.* A centrosomal Cdc20-APC pathway controls dendrite morphogenesis in postmitotic neurons. *Cell* **136**, 322–336 (2009).
39. Songyang, Z. *et al.* A structural basis for substrate specificities of protein Ser/Thr kinases: primary sequence preference of casein kinases I and II, NIMA, phosphorylase kinase, calmodulin-dependent kinase II, CDK5, and Erk1. *Mol. Cell. Biol.* **16**, 6486–6493 (1996).
40. Zur, A. & Brandeis, M. Securin degradation is mediated by fzy and fzr, and is required for complete chromatid separation but not for cytokinesis. *EMBO J.* **20**, 792–801 (2001).
41. Brocke, L., Srinivasan, M. & Schulman, H. Developmental and regional expression of multifunctional Ca^{2+}/calmodulin-dependent protein kinase isoforms in rat brain. *J. Neurosci.* **15**, 6797–6808 (1995).
42. Higginbotham, H.R. & Gleeson, J.G. The centrosome in neuronal development. *Trends Neurosci.* **30**, 276–283 (2007).
43. Badano, J.L., Teslovich, T.M. & Katsanis, N. The centrosome in human genetic disease. *Nat. Rev. Genet.* **6**, 194–205 (2005).
44. Dammermann, A. & Merdes, A. Assembly of centrosomal proteins and microtubule organization depends on PCM-1. *J. Cell Biol.* **159**, 255–266 (2002).
45. Dierssen, M. & Ramakers, G.J. Dendritic pathology in mental retardation: from molecular genetics to neurobiology. *Genes Brain Behav.* **5** (suppl. 2): 48–60 (2006).
46. Pardo, C.A. & Eberhart, C.G. The neurobiology of autism. *Brain Pathol.* **17**, 434–447 (2007).

ONLINE METHODS

Antibodies. The antibodies purchased for this research include monoclonal CaMKIIβ and CaMKIIα antibodies (Zymed), polyclonal rabbit phosphoThr287-CaMKIIβ antibody (Badrilla), polyclonal rabbit CaMKIIα, CaMKIIβ and ERK1/2 antibodies (Cell Signaling), rabbit Bad, mouse Myc, mouse Actin, rabbit Cdc20, rabbit Id1 and rabbit SnoN antibodies (Santa Cruz Biotechnology), mouse Flag, γ-tubulin, calbindin and microtubule associated protein 2 antibodies (Sigma), mouse α-tubulin, mouse hemagglutinin (clone 16B12) and rabbit pericentrin antibodies (Covance), rabbit and mouse GFP antibodies (Invitrogen), mouse β-galactosidase antibody (Promega) and rat hemagglutinin (clone 3F10) antibody (Roche). Rabbit 14-3-3ε antibody was a generous gift from Alastair Aitken (University of Edinburgh). Rabbit PCM1 antibody has been described[44]. Rabbit phosphoSer51-Cdc20 antibody was generated by injection of rabbits with the phospho-peptide CAANRSHpSAGRTPG (amino acids 44–57 in rat Cdc20) and purified by standard affinity purification at Cell Signaling Technology.

Plasmids. Rat CaMKIIβ and CaMKIIα cDNA was a gift from Tobias Meyer (Stanford University) and was cloned into pcDNA3 to produce Myc epitope-tagged and Flag epitope-tagged CaMKIIβ expression plasmids and into pEGFP-C1 (Clontech) to produce an N-terminal GFP-tagged CaMKIIβ expression plasmid. C-terminal GFP-tagged chicken PCM1 in pEGFP-N1 (Clontech) has been described[44]. GFP-centrin was a gift from Karl Munger (Harvard University).

shRNA plasmids were produced by insertion of the following oligonucleotides into pBS-U6 or pBS-U6-CMV-GFP: U6-CaMKIIβi.1, 5′-GTCCGACGCTGTG TCAAGCACAAGTTAACAGCTTGACACAGCGTCGGACCTTTTTG-3′; U6-CaMKIIβi.2, 5′-GCAGCTAAGATCATTAACACGCAAGTTAACGGTGTT AATGATCTTAGCTGCCTTTTTG-3′; U6-PCM1i.1, 5′-CTTGAAGCTCTAAT GGCTGAACTTTTTG-3′; U6-PCM1i.2, GGAGCACATGGATGAAGTATGCT TTTTG-3′. U6-Cdc20i and U6-Id1i have been described[38].

Mutations in CaMKIIβ and other expression plasmids were performed using standard protocols and confirmed by sequencing. FABD and C-terminal variable region deletion constructs were generated by deletion of the FABD and C-terminal variable region (consisting of the linker and C-terminal segment) as described[36,41]. CaMKIIβ-RESΔAssoc and CaMKIIβ-RESΔCTS/Assoc were created by standard PCR subcloning of CaMKIIβ lacking the association domain and CaMKIIβ lacking the association domain and C-terminal variable region[36,41] into expression constructs.

Primary neuron cultures and transfection. Primary cerebellar granule neurons were prepared from P6 rat pups and transfected via a modified calcium phosphate protocol as described[28]. To avoid the possibility that morphological effects of RNAi or protein expression were a result of changes in cell survival, we included the expression plasmid encoding the anti-apoptotic protein Bcl-XL in all neuronal transfections. The CaMKIIβ RNAi-induced dendrite phenotype was identical in the presence or absence of Bcl-XL expression (data not shown). Hippocampal and cerebral cortical neuron cultures were prepared from embryonic day 18 (E18) rat embryos as described[47] and transfected using a modified calcium phosphate protocol.

Cerebellar slice cultures and *in vivo* electroporation. P6 and P10 rat cerebella were prepared as described[27]. Individual neurons in P6 and P10 slices were transfected after 2 or 4 d, respectively, using biolistics (Helios gene gun, Bio-Rad) as described[27]. Four days after transfection, slices were subjected to immunohistochemical analyses.

All experiments using live animals were approved by the Harvard Medical School Standing Committee on Animals and strictly conform to their regulatory standards. *In vivo* electroporation of P3 Sprague–Dawley rat pups was performed as described[28]. Five or nine days after electroporation (P8 or P12, respectively), rats were killed and cerebella collected. Coronal sections of cerebella (40 μm) were prepared and subjected to immunohistochemistry with the GFP and calbindin antibody and the DNA dye bisbenzimide (Hoechst 33258).

Time-lapse analyses of dendrite morphogenesis. For initial time-lapse analyses, granule neurons were plated on etched coverslips (Bellco) and transfected after 1 d *in vitro* (DIV1). Beginning at DIV2, individual neurons were monitored by noting their position on the grid and obtaining sequential images of that position over the 48 h period of analysis. Neurons were randomly selected. For CaMKIIβ knockdown experiments, granule neurons were similarly plated but transfected on DIV0. Beginning at DIV2, CaMKIIβ knockdown and control U6-transfected neurons were imaged.

For live confocal imaging analyses, granule neurons were plated on glass-bottom multi-well plates (MatTek) and transfected as described for initial time-lapse analyses. Beginning at DIV2, live imaging was performed using a PerkinElmer Life and Analytical Sciences UltraVIEW spinning-disk confocal system with a Nikon Ti-E Perfect-Focus microscope. An environment-controlled chamber maintained the neurons at 37 °C, 5% CO_2. Images were acquired and analyzed using Volocity software (PerkinElmer Life and Analytical Sciences). A 20× objective was used in combination with an automated stage to capture images across 48 h. Neurons were chosen randomly and followed by saving their coordinates on the motorized stage. Dendrite extension and retraction events were defined as events where dendrites extended or retracted by greater than 2 μm over the 1 h period analyzed.

Immunocytochemistry. For visualization of centrosomal proteins, neurons were fixed in absolute methanol for 10 min at −20 °C and subjected to immunofluorescence analysis after blocking and staining with the indicated antibodies according to standard protocols. For other immunocytochemistry experiments, neurons were fixed in 4% (vol/vol) paraformaldehyde for 20 min at 25 °C and analyzed as described[28]. Cells were counted as having dispersed Cdc20 if there were three or more discrete Cdc20 puncta or if puncta of Cdc20 immunofluorescence localized away from the centrosome as based on GFP-centrin or endogenous pericentrin.

Immuno-electron microscopy analyses. Granule neurons were fixed in 4% paraformaldehyde with 0.025% glutaraldehyde and 5 μg ml^{-1} Taxol in Brinkley buffer 1980 (80 mM PIPES, pH 6.8, 1 mM $MgCl_2$, 1 mM EGTA) for 10 min at 37 °C. Neurons were permeabilized in 0.1% Triton X-100 in BRB80 with 5 μg ml^{-1} taxol and immunostained with CaMKIIβ or control (IgG) antibody overnight. Samples were then incubated with 5 nm gold-conjugated Protein A, sectioned using an ultramicrotome (Reichert), and collected on coated copper grids. Sections were visualized using a JEOL 1200EX transmission electron microscope.

Centrosomal fractionation of primary neuronal lysates. Centrosomal fractions from granule neurons or cortical neurons were isolated as described[48]. Briefly, neurons were treated with 2 μg ml^{-1} nocodazole and 1 μg ml^{-1} cytochalasin D at 37 °C for 1 h, washed sequentially in cold PBS, 0.1× PBS + 8% sucrose (wt/wt), and 8% sucrose, and then lysed in a hypotonic lysis buffer with a protease inhibitor cocktail (leupeptin, aprotinin, phenylmethylsulfonyl fluoride and pepstatin) and NaF. Homogenates were centrifuged at 1,500*g* for 10 min, and PIPES, pH 7.2 (final concentration 10 mM) was added to the postnuclear supernatant. The postnuclear fraction was treated with 2 μg ml^{-1} DNase I for 30 min at 4 °C and then layered on top of a discontinuous sucrose gradient (70%/50%/40% sucrose (wt/wt) in 10 mM PIPES pH 7.2, 0.1% Triton X-100 and 0.1% β-mercaptoethanol) and subjected to ultracentrifugation at 20,000 r.p.m. in a SW-60 rotor (Beckman) for 1.5 h. Fractions were collected from the bottom of the tube and analyzed by immunoblotting.

Immunoprecipitation analyses. Cells were lysed in 150 mM NaCl, 20 mM Tris-HCl pH 7.5, 1 mM ethylenediaminetetraacetic acid (EDTA), 1% nonyl phenoxypolyethoxylethanol 40 (NP40) containing protease inhibitors. Lysates were briefly precleared with a combination of Protein A and G Sepharose beads (GE Healthcare) and then incubated with either the appropriate antibody or antibody-conjugated beads overnight. For non-conjugated antibodies, the antibody–protein complexes were immunoprecipitated with Protein A/G beads. Immunoprecipitated proteins bound to beads were washed several times and lysates were analyzed by SDS-PAGE and transferred to a nitrocellulose membrane for immunoblotting analysis.

Purification of GST-Cdc20 fusion proteins and *in vitro* kinase assays. GST-Cdc20 constructs were expressed in *Escherichia coli* and purified by solubilization in buffer containing 500 mM NaCl, 20 mM Tris pH 7.5, 0.2 mM EGTA, 0.2 mM EDTA and protease inhibitors. Purified GST-Cdc20 proteins were incubated with CaMKII purified from rat forebrain (Calbiochem) for 10 min at 37 °C in the presence of 10 mM HEPES pH 7.5, 10 mM $MgCl_2$, 1 mM Na_3VO_4, 0.5 mM $CaCl_2$,

doi:10.1038/nn.2857

2.5 mM dithiothreitol, 10 μg ml^{-1} calmodulin and 200 μM ATP (containing trace γ-^{32}P-ATP). Reactions were terminated by addition of sodium dodecyl sulfate (SDS) sample buffer, and samples were resolved by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and stained with Coomassie brilliant blue. Radioactivity was visualized by Phosphorimager (Amersham Biosciences) analysis.

Mass spectrometry analyses. The Coomassie-stained band corresponding to the putative Cdc20 protein was excised from the gel and destained using 50% CH_3CN (vol/vol), 50 mM NH_4HCO_3. The gel was dehydrated with 100% CH_3CN, followed by rehydration with 2 M urea in 25 mM Tris buffer, pH 8.8, containing 25 ng μl^{-1} of the endopeptidase Lys-C. In-gel digestion was performed overnight at 37 °C. Liquid chromatography-mass spectrometry (LC-MS/MS) data were obtained using an LTQ-Orbitrap XL hybrid mass spectrometer (Thermo Fisher). The sample was loaded onto a pulled fused silica microcapillary column (125 μm, 15 cm) packed with C_{18} reverse-phase resin (Magic C18AQ; Michrom Bioresources). Peptides were separated using an Agilent 1200 series binary pump across a 50-min linear gradient of 4–25% CH_3CN in 0.125% HCOOH (vol/vol) at a flow rate of 800 nl min^{-1}. In each data collection cycle, one full MS scan (300–1,500 *m/z*) was acquired in the Orbitrap (6×10^4 resolution, automatic gain control (AGC) target of 10^6), followed by ten data-dependent MS/MS scans in the LTQ (AGC target, 10^4; threshold 5×10^3) using the ten most abundant ions and electron transfer dissociation for fragmentation. We used a reaction time of 69 ms and an anion target value of 2×10^5 reagent ions. SEQUEST analysis (v. 27, rev. 12, ThermoFinnigan) was used to search MS/MS spectra for rat Cdc20 sequence to identify Cdc20 phosphopeptides. The search parameters used for post-translational modification included 79.96633 Da on serine, threonine and tyrosine for phosphorylation.

Luciferase reporter assays. The Myc-PACT-securin-luciferase or Myc-PACT-securin DBM-luciferase plasmids were transfected with the indicated RNAi plasmid or control U6 vector in granule neurons as described[38]. Four days later, neurons were treated with cycloheximide (100 μg ml^{-1}) for 4 h and luciferase in lysates measured using luminometry (Promega). Luminescence values obtained with securin-luciferase transfection were divided by those of securin DBM-luciferase (Centrosomal Securin/Securin DBM), yielding a value inversely proportional to centrosomal Cdc20-APC activity.

Analysis of neuronal morphology and imaging. To analyze the axonal and dendritic morphology of primary neurons in culture, in slices and *in vivo*, images of individual neurons were captured randomly in a blinded manner on a Nikon Eclipse TE2000 epifluorescence microscope using a digital CCD camera (Diagnostic Instruments). SPOT software was used to measure individual process length by tracing. Axons and dendrites were identified in transfected neurons based on morphology and selective expression of MAP2 and Tau1 in dendrites and axons, respectively (data not shown). Total dendrite length was determined by summing the lengths of all dendrite processes measured from a single neuron. To analyze dendrite morphology *in vivo*, granule neurons residing in the IGL were selected for morphometry. Percent parallel fiber association was determined by counting the number of parallel fibers and cell bodies present in a specific region of a section in consecutive sections in a blinded manner as described[49]. To analyze cell survival, neurons were transfected with the β-galactosidase expression plasmid and subjected to immunocytochemistry using the β-galactosidase antibody and the DNA dye bisbenzimide (Hoechst 33258). Cell survival was scored by assessment of process fragmentation and nuclear condensation.

Confocal images were collected using an Olympus IX81 microscope with FluoView1000 scanning confocal unit (taken with a 40×, 0.90 numerical aperture Olympus UPlanSApo or 60×, 1.42 numerical aperture oil Olympus PlanApoN objective). Labeled neurons were excited at 405 nm, 488 nm and 559 nm, and emission was collected at 425–475 nm, 500–545 nm and 575–675 nm for Hoechst, Alexa Fluor-488 and Cy3, respectively.

Line scan analyses. Images taken from transfected neuron cohorts using identical acquisition parameters were subjected to line scan analyses. The immunofluorescent signal was analyzed using Olympus Fluoview-1000 image analysis software by taking pericentrin signal as the location of the centrosome and tracing GFP-Cdc20 signal and quantifying immunofluorescence intensity along the line. Line scan intensity plots from individual neurons were aligned at the centrosomal peak (defined as zero on the *x* axis) and were combined to generate an average for the cohort. Plots were thresholded at the level of background signal. Intensity values along the traced line are reported as a percentage of the maximum centrosomal peak signal.

Statistics. All analyses were completed from a minimum of three independent experiments. The number of cells contributing to each condition was equally distributed across independent experiments. Statistical analyses were performed with GraphPad Prism 4.0. All histograms are presented as mean + s.e.m. unless otherwise noted. Student's *t*-test was used for comparisons in experiments with two sample groups. In experiments with more than two sample groups, ANOVA was performed followed by Bonferroni's *post hoc* test. For nonparametric data with more than two sample groups, the Kruskal–Wallis test was used.

47. Goslin, K., Asmussen, H. & Banker, G.A. *Rat Hippocampal Neurons in Low-Density Culture* (MIT Press, Cambridge, Massachusetts, USA, 1998).
48. Bornens, M., Paintrand, M., Berges, J., Marty, M.C. & Karsenti, E. Structural and chemical characterization of isolated centrosomes. *Cell Motil. Cytoskeleton* **8**, 238–249 (1987).
49. Stegmüller, J. *et al.* Cell-intrinsic regulation of axonal morphogenesis by the Cdh1-APC target SnoN. *Neuron* **50**, 389–400 (2006).

Appendix C

C. Elegans SIRT6/7 Homolog SIR-2.4 Promotes DAF-16 Relocalization and Function During Stress

Attributions:

- This appendix contains work published as Chiang, W. C., Tishkoff, D. X., Yang, B., Wilson-Grady, J., Yu, X., Mazer, T., Eckersdorff, M., Gygi, S. P., Lombard, D. B., and Hsu, A. L., C. elegans SIRT6/7 homolog SIR-2.4 promotes DAF-16 relocalization and function during stress. PLoS Genet 2012, 8, (9), e1002948.

- J.T Wilson-Grady performed the LC-MS/MS analysis which identified DAF-16 acetylation sites, including database searching and site localization, summarized in table S3.

- S.P. Gygi provided instrumentation and computational tools for LC-MS/MS, and the relevant data analysis.

- D. B. Lombard and A. L. Hsu designed the experiments and prepared the manuscript

- All other authors performed the remaining experiments and associated data analysis, where W. C. Chiang, D. X. Tishkoff and B. Yang contributed the majority equally.

OPEN ACCESS Freely available online

PLOS | GENETICS

C. elegans SIRT6/7 Homolog SIR-2.4 Promotes DAF-16 Relocalization and Function during Stress

Wei-Chung Chiang[1☯], Daniel X. Tishkoff[2☯], Bo Yang[2☯], Joshua Wilson-Grady[3], Xiaokun Yu[4], Travis Mazer[4], Mark Eckersdorff[2¤], Steven P. Gygi[3], David B. Lombard[2,5]*, Ao-Lin Hsu[1,4,5]*

1 Department of Molecular and Integrative Physiology, University of Michigan, Ann Arbor, Michigan, United States of America, **2** Department of Pathology, University of Michigan, Ann Arbor, Michigan, United States of America, **3** Department of Cell Biology, Harvard Medical School, Boston, Massachusetts, United States of America, **4** Department of Internal Medicine, Division of Geriatric Medicine, University of Michigan, Ann Arbor, Michigan, United States of America, **5** Institute of Gerontology and the Geriatrics Center, University of Michigan, Ann Arbor, Michigan, United States of America

Abstract

FoxO transcription factors and sirtuin family deacetylases regulate diverse biological processes, including stress responses and longevity. Here we show that the *Caenorhabditis elegans* sirtuin SIR-2.4—homolog of mammalian SIRT6 and SIRT7 proteins—promotes DAF-16–dependent transcription and stress-induced DAF-16 nuclear localization. SIR-2.4 is required for resistance to multiple stressors: heat shock, oxidative insult, and proteotoxicity. By contrast, SIR-2.4 is largely dispensable for DAF-16 nuclear localization and function in response to reduced insulin/IGF-1-like signaling. Although acetylation is known to regulate localization and activity of mammalian FoxO proteins, this modification has not been previously described on DAF-16. We find that DAF-16 is hyperacetylated in *sir-2.4* mutants. Conversely, DAF-16 is acetylated by the acetyltransferase CBP-1, and DAF-16 is hypoacetylated and constitutively nuclear in response to *cbp-1* inhibition. Surprisingly, a SIR-2.4 catalytic mutant efficiently rescues the DAF-16 localization defect in *sir-2.4* null animals. Acetylation of DAF-16 by CBP-1 *in vitro* is inhibited by either wild-type or mutant SIR-2.4, suggesting that SIR-2.4 regulates DAF-16 acetylation indirectly, by preventing CBP-1-mediated acetylation under stress conditions. Taken together, our results identify SIR-2.4 as a critical regulator of DAF-16 specifically in the context of stress responses. Furthermore, they reveal a novel role for acetylation, modulated by the antagonistic activities of CBP-1 and SIR-2.4, in modulating DAF-16 localization and function.

Citation: Chiang W-C, Tishkoff DX, Yang B, Wilson-Grady J, Yu X, et al. (2012) *C. elegans* SIRT6/7 Homolog SIR-2.4 Promotes DAF-16 Relocalization and Function during Stress. PLoS Genet 8(9): e1002948. doi:10.1371/journal.pgen.1002948

Editor: Kaveh Ashrafi, University of California San Francisco, United States of America

Received May 21, 2011; **Accepted** July 27, 2012; **Published** September 13, 2012

Funding: This work was supported by R01 grants from NIA (AG028516; A-LH) and NIGMS (GM101171; DBL), a research grant from the American Federation for Aging Research (DBL), a Rackham Predoctoral Fellowship (W-CC), an NIA training grant (AG000114; DXT), and a pilot award from the Nathan Shock Center at the University of Michigan (P30 AG013283; DBL). DBL is a New Scholar in Aging of the Ellison Medical Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: davidlom@umich.edu (DBL); aolinshu@umich.edu (A-LH)

¤ Current address: Regeneron Pharmaceuticals, Tarrytown, New York, United States of America

☯ These authors contributed equally to this work.

Introduction

Elucidation of mechanisms regulating stress resistance and longevity has been aided tremendously by the use of invertebrate models. FoxO transcription factors regulate multiple biological processes in many organisms [1]. In *C. elegans* and *Drosophila*, increased FoxO activity promotes longevity, fat storage, and stress resistance [2]. Mammals possess four FoxO homologs, with partially redundant and distinct functions: FoxO1, FoxO3A, FoxO4, and FoxO6 [1]. These proteins regulate apoptosis, cell cycle arrest, oxidative defense, DNA repair, metabolism, differentiation, stem cell function, and tumor suppression in a cell type- and context-specific manner [1].

FoxO activity is tightly controlled, and subcellular localization is a principal mechanism of FoxO regulation [3]. In this context, insulin/IGF-1-like signaling (IIS) is the major influence on FoxO function. IIS leads to FoxO phosphorylation and cytoplasmic segregation in a complex with 14-3-3 chaperone proteins. Conversely, stress stimuli promote nuclear translocation of FoxO proteins by multiple mechanisms, including activation of stress kinases that modify FoxO proteins on residues distinct from those phosphorylated in IIS [2]. In response to oxidative insult, mammalian FoxO proteins are acetylated [4,5,6,7], mono-ubiquitylated [8], and phosphorylated [9,10,11,12]. Overall, these post-translation modifications function as a "FoxO code", providing a means by which FoxO activity is finely regulated in response to various stimuli to promote altered metabolism, stress responses, or cell death [3].

The sirtuins are an evolutionarily conserved protein family impacting many biological processes, including longevity, stress responses, metabolism, and cancer [13]. Sirtuins modify target proteins by means of their NAD^+-dependent lysine deacetylase and ADP-ribosyltransferase activities. Lysine acetylation has emerged as a post-translational modification with a key role in modulating protein function, akin to phosphorylation [14]. Mammals possess seven sirtuins, SIRT1-SIRT7. The *C. elegans* genome encodes four sirtuins, SIR-2.1 through SIR-2.4, corresponding to mammalian SIRT1 (SIR-2.1), SIRT4 (SIR-2.2 and SIR-2.3), and SIRT6/7 (SIR-2.4) [15]. SIR-2.1 has been

Author Summary

Sensing and responding appropriately to environmental insults is a challenge facing all organisms. In the roundworm *C. elegans*, the FoxO protein DAF-16 moves to the nucleus in response to stress, where it regulates gene expression and plays a key role in ensuring organismal survival. In this manuscript, we characterize SIR-2.4 as a novel factor that promotes DAF-16 function during stress. SIR-2.4 is a member of a family of proteins called sirtuins, some of which promote increased lifespan in model organisms. Worms lacking SIR-2.4 show impaired DAF-16 nuclear recruitment, DAF-16–dependent gene expression, and survival in response to a variety of stressors. SIR-2.4 regulates DAF-16 by indirectly affecting levels of a modification called acetylation on DAF-16. Overall, our work has revealed SIR-2.4 to be a key new factor in stress resistance and DAF-16 regulation in *C. elegans*. Future studies will address whether mammalian SIR-2.4 homologs SIRT6 and SIRT7 act similarly towards mammalian FoxO proteins.

implicated in numerous physiologic processes, including stress responses [16,17,18,19,20,21,22,23]. In contrast, functions of other worm sirtuins are largely uncharacterized [24].

In different organisms, sirtuins modulate FoxO activity via diverse means. In *C. elegans*, there are reports that SIR-2.1 extends longevity in a DAF-16-dependent manner [21,25,26,27]. However, this is currently a disputed finding [28,29]. In mammals, SIRT1 directly deacetylates FoxO proteins in response to oxidative stress. The effect of FoxO deacetylation is somewhat controversial [4,6,7,30,31], but it is likely that the overall outcome is to promote DNA repair and cell cycle arrest while inhibiting apoptosis [3]. SIRT2 also deacetylates FoxO proteins to inhibit adipocytic differentiation [32,33] and regulate levels of intracellular reactive oxygen species (ROS) [34]. SIRT1 and SIRT2-mediated deacetylation of FoxO1 promotes nuclear accumulation of this protein [5,32]. Acetylation of FoxO1 has also been reported to attenuate DNA binding, to promote AKT-mediated FoxO1 phosphorylation, and to direct FoxO1 to nuclear PML bodies [7,35]. The mitochondrial sirtuin SIRT3 has also been proposed to modulate FoxO function [36,37].

Here, we characterize functions of the *C. elegans* sirtuin SIR-2.4, a protein about which little is currently known. We find that SIR-2.4 plays a crucial role in promoting DAF-16 transcriptional activity and stress-induced nuclear localization, and is required for normal stress resistance in the worm. However, SIR-2.4 is largely dispensable for DAF-16 nuclear localization and function in the context of reduced IIS. We show directly for the first time that DAF-16 itself is acetylated, by the acetyltransferase CBP-1. SIR-2.4 attenuates DAF-16 acetylation in a non-catalytic activity-dependent manner, and catalytic function of SIR-2.4 is dispensable for regulation of DAF-16 localization in response to stress. Acetylation of DAF-16 by CBP-1 is inhibited in the presence of SIR-2.4. Our results indicate acetylation plays a key role in regulating DAF-16 localization and function, and that levels of this modification are modulated by the antagonistic functions of CBP-1 and SIR-2.4.

Results

SIR-2.4 is required for efficient stress-induced DAF-16 nuclear localization

Mammalian SIRT6 and SIRT7 proteins both promote genotoxic stress resistance [38,39]. We therefore tested a potential role for SIR-2.4 in stress resistance and DAF-16 regulation. We generated an RNAi construct encoding nucleotides 1 467 of the *SIR-2.4* open reading frame in the RNAi vector L4440. *sir-2.4* knockdown (KD) resulted in no obvious defects under basal conditions. However, *sir-2.4* RNAi severely impaired stress-induced DAF-16 nuclear localization (Figure 1A). *sir-2.4* RNAi inhibited DAF-16 nuclear translocation in response to either heat shock or oxidative insult by ~50% shortly after stress induction (Figure 1B). At later timepoints, DAF-16 did translocate to the nucleus in *sir-2.4* KD worms (see below and data not shown).

sir-2.4 deletion is reported to confer lethality/sterility (National Bioresource Project, Japan). However, during the course of analyzing the effects of *sir-2.4* RNAi, we obtained a viable strain with a deletion removing all but the initial 9 amino acids of the *SIR-2.4* open reading frame (kind gift of H.R. Horvitz; see *Materials and Methods* section for a complete description of this strain). As with *sir-2.4* RNAi, *sir-2.4* KO animals showed no apparent defects under unstressed conditions. However, like *sir-2.4* KD worms, *sir-2.4* knockouts (KOs) showed significantly delayed stress-induced DAF-16 nuclear translocation in response to oxidative stress (Figure 1C; $p<0.001$ by Poisson regression analysis) and heat shock (Figure 1D; $p<0.001$). We conclude that SIR-2.4 is dispensable for viability and fertility, but plays a crucial role in directing DAF-16 to the nucleus in response to stress, particularly at early time points following stress induction.

It has been reported that SIR-2.1 and 14-3-3 proteins act in concert to activate DAF-16 [21,26]. To examine whether SIR-2.1 plays a role in directing DAF-16 to the nucleus in response to stress, we assessed the effect of *sir-2.1* KD on DAF-16 nuclear localization. *sir-2.1* KD alone had little impact on DAF-16 nuclear recruitment in response to either oxidative insult (Figure 1C; $p<0.72$) or heat stress (Figure 1D; $p<0.44$), indicating that SIR-2.1 does not play a major role in stress-induced DAF-16 nuclear localization, consistent with published data [17]. Moreover, KD of *sir-2.1* in the context of *sir-2.4* mutation did not produce any statistically significant additional delay in DAF-16 nuclear recruitment versus *sir-2.4* KO alone (Figure 1C 1D; $p<0.89$ and $p<0.13$ for oxidative and heat stress, respectively). We conclude that SIR-2.4, but not SIR-2.1, plays a major role in promoting rapid nuclear recruitment of DAF-16 in response to oxidative stress or heat shock. These results imply that SIR-2.1 and SIR-2.4 act in distinct pathways to influence DAF-16 functions.

While stress-induced subcellular translocation of DAF-16 was affected in all cell types by *sir-2.4* inhibition, we did note that *sir-2.4* KD and KO had a bigger impact in head hypodermis cells than in intestinal cells (data not shown). SIR-2.4 was very weakly expressed in most cell types, but showed much stronger expression in a subset of head and tail neurons as well as in a subset of somatic gonad cells (Figure S1). We have not yet formally assessed the tissue requirements for SIR-2.4 function in the regulation of DAF-16 localization, though our functional data suggest a cell autonomous mechanism (see below).

SIR-2.4 promotes DAF-16–dependent transcription under basal and stress conditions

DAF-16 carries out its functions by transcriptional regulation of a large number of target genes [40,41,42]. The role of SIR-2.4 in DAF-16-dependent gene expression was tested in the context of six well-known DAF-16 targets. We confirmed the published role of DAF-16 in regulating expression of all six of these genes (Figure S2A). Under both basal and stress conditions, *sir-2.4* RNAi led to decreased mRNA levels of three genes positively regulated by DAF-16 (*SOD-3*, *HSP-16.1*, and *DOD-3*), (Figure 2, top row).

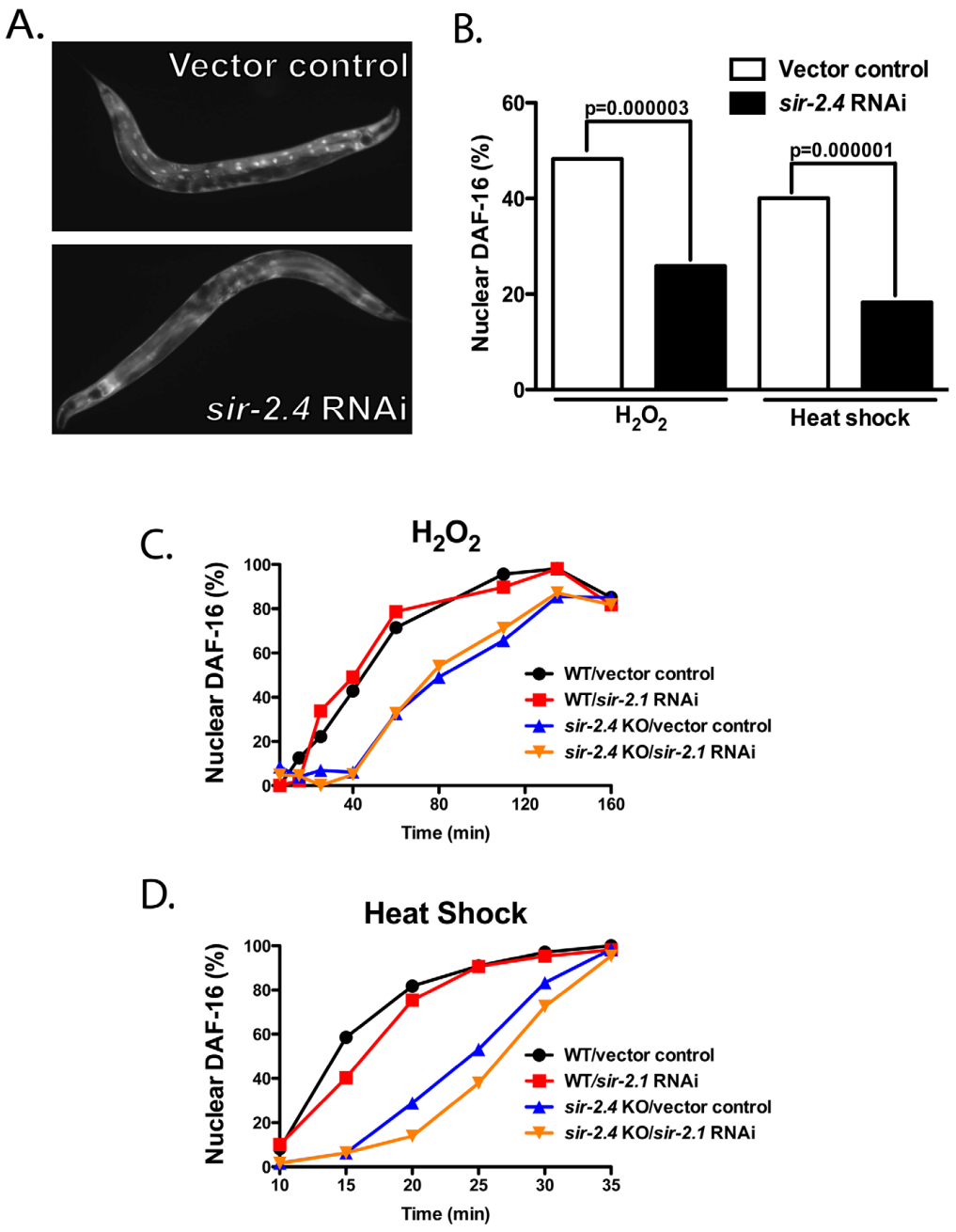


Figure 1. SIR-2.4, but not SIR-2.1, is required for stress-induced DAF-16 nuclear localization. TJ356 animals carrying an integrated *daf-16::gfp* array were fed either vector control or *sir-2.4* RNAi bacteria for at least one generation before being subjected to heat-shock or oxidative stress. (A) Images of TJ356 animals grown on control or *sir-2.4* RNAi bacteria after 15 min heat-shock. (B) Quantification of DAF-16::GFP nuclear

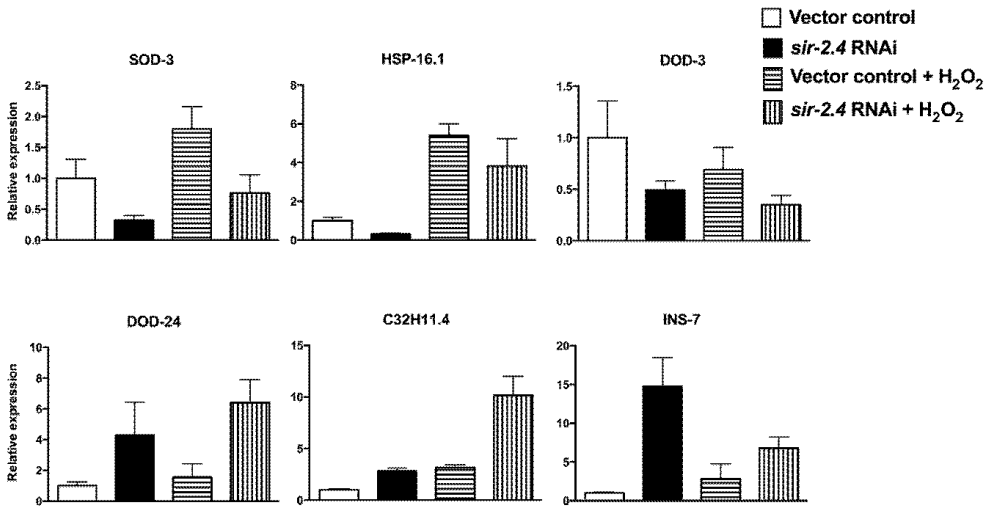
accumulation in response to heat-shock (35°C for 15 min.) or oxidative stress (1.5 mM H_2O_2 for 1 hr). Worms were scored for the presence or absence of GFP accumulation within the intestinal nuclei (n = 120 or greater for all treatments). An animal was scored as having nuclear GFP if one or more intestinal nuclei contained DAF-16-GFP. (C–D) Time course analysis of DAF-16::GFP nuclear accumulation in response to stress. TJ356 or EQ200 [*sir-2.4(n5137); daf-16::gfp*] animals grown on either control or *sir-2.1* RNAi bacteria were subjected to (C) heat-shock (35°C) or (D) oxidative stress (1.5 mM H_2O_2). Worms were scored for GFP accumulation within the head hypodermic nuclei at day 1 of adulthood (n = 30~50) every 5–30 min.
doi:10.1371/journal.pgen.1002948.g001

Expression of these DAF-16 targets was similarly attenuated in the *sir-2.4* KO strain (Figure S2B). Conversely, expression of three genes negatively regulated by DAF-16 was greatly increased by *sir-2.4* RNAi (*DOD-24*, *C32H11.4*, and *INS-7*) (Figure 2, bottom row). We conclude that SIR-2.4 promotes DAF-16 transcriptional function under both basal and oxidative stress conditions.

SIR-2.4 is required for normal stress resistance

DAF-16 is a key regulator of stress responses in *C. elegans* [43]. We therefore tested the impact of SIR-2.4 on stress resistance. *sir-2.4* KO and *sir-2.4* KD worms were hypersensitive to heat shock (Figure 3A, Table S1) and oxidative insult (Figure 3B, Table S1). Simultaneous inhibition of both *sir-2.4* and *daf-16* increased stress sensitivity to a similar extent as observed in *daf-16* single mutants, and the degree of hypersensitivity conferred by either single KO/KD alone was similar (Figure 3A 3B, Figure S2C S2D), suggesting that SIR-2.4 and DAF-16 modulate stress resistance via a common pathway. Conversely, overexpression of SIR-2.4 did not produce increased stress resistance (Figure S3A S3B; Table S1); hence SIR-2.4 levels are not limiting for DAF-16 regulation and stress resistance.

Expression of fluorescently tagged polyglutamine (polyQ) repeat-containing proteins in *C. elegans* body wall muscle causes paralysis that is antagonized by DAF-16 [44,45], a model for proteotoxicity occurring in human neurodegenerative diseases such as Huntington's disease. The role of SIR-2.4 in proteotoxicity resistance was assessed. Worms expressing 35 glutamine residues conjugated to YFP in body wall muscle (*unc-54p::Q35::YFP*) were



P values	*sod-3*	*hsp-16.1*	*dod-3*	*dod-24*	*C32H11.4*	*ins-7*
Vector control vs *sir-2.4* RNAi	**0.018**	**0.017**	**0.012**	**0.011**	**0.003**	**0.023**
Vector control vs Vector control + H_2O_2	**0.016**	**0.004**	0.280	0.377	**0.002**	0.243
Vector control + H_2O_2 vs *sir-2.4* RNAi + H_2O_2	**0.005**	**0.018**	0.096	**0.014**	**0.020**	**0.050**
sir-2.4 RNAi vs *sir-2.4* RNAi + H_2O_2	0.054	**0.050**	0.121	0.238	**0.018**	0.051

Figure 2. SIR-2.4 is required for optimal DAF-16-dependent gene expression. Wild-type N2 animals fed on either vector control or *sir-2.4* RNAi bacteria from the time of hatching were exposed to 10 mM H_2O_2 for 80 min. Relative mRNA levels of SOD-3, HSP-16.1, DOD-3, DOD-24, C32H11.4, and INS-7 were measured by quantitative RT-PCR and the means of three different sample sets are shown. Relative mRNA levels were normalized against ACT-1 (beta-actin). Error bars: ± STD. Statistical significance as determined by two-tailed t-test is shown in the table below; significant differences are represented in black font.
doi:10.1371/journal.pgen.1002948.g002

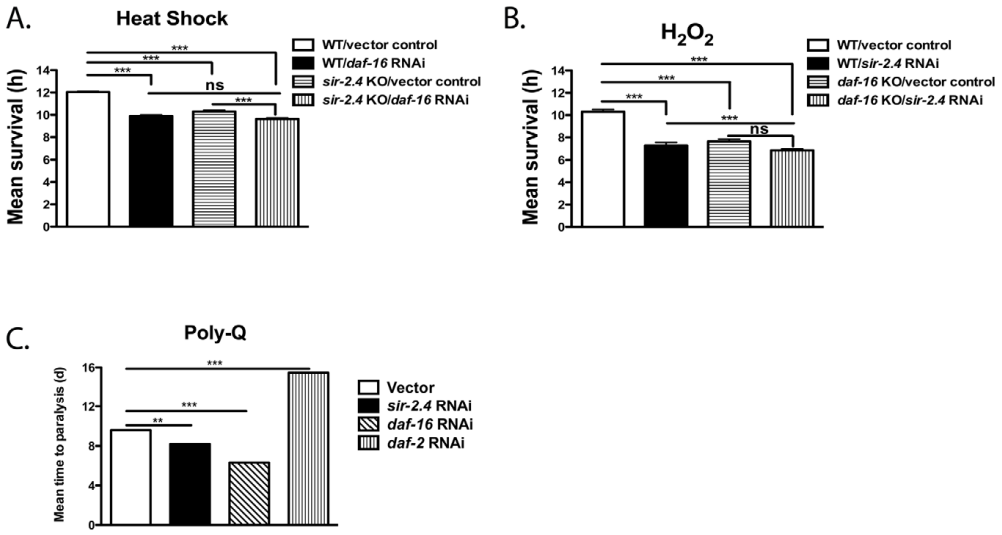


Figure 3. SIR-2.4 promotes stress resistance. (A) Wild-type N2 or *sir-2.4(n5137)* worms grown on vector control or *daf-16* RNAi bacteria were subjected to heat-shock at 35°C and scored for viability every 1-2 hours. (B) Wild-type N2 or *daf-16(mu86)* worms grown on vector control or *sir-2.4* RNAi bacteria were treated with 1.5 mM H_2O_2 and scored for viability every 1–2 hours. Data are mean survival±SEM, in hours for (A–B). ***, $p<0.001$; ns, $p>0.05$. See Table S1 for statistical analysis. (C) AM140 worms expressing a poly-Q tract (Q35::YFP) were seeded on the RNAi bacteria indicated and scored for poly-Q induced paralysis every other day. Data are mean time to paralysis in days±SEM. ***, $p<0.001$; **, $p<0.01$; ns, $p>0.05$.
doi:10.1371/journal.pgen.1002948.g003

grown on *daf-2* RNAi, *daf-16* RNAi, *sir-2.4* RNAi, or control bacteria. *daf-2* RNAi delayed, and both *daf-16* and *sir-2.4* RNAi accelerated, the onset of Q35::YFP-induced paralysis (Figure 3C). Thus, like DAF-16, SIR-2.4 is required for resistance to multiple stressors, including proteotoxic injury.

SIR-2.4 is largely dispensable for DAF-16 function in response to reduced IIS

Reduced IIS promotes DAF-16 nuclear localization and functions independently of exogenous stress. We performed several assays to determine whether SIR-2.4 regulates DAF-16 in the context of IIS. Many manipulations that reduce IIS increase lifespan in *C. elegans*. However, neither *sir-2.4* RNAi nor *SIR-2.4* overexpression affected the lifespan of wild-type worms (Figure 4A and Figure S3C, Table S2), nor did *sir-2.4* RNAi suppress increased longevity of *daf-2 (e1370)* insulin/IGF-I-like receptor mutants (Figure 4A, Table S2). *sir-2.4* deletion or *sir-2.4* RNAi minimally impacted DAF-16 nuclear translocation induced by reduced IIS (Figure 4B and Figure S2E), in contrast to its potent impact on stress-induced DAF-16 relocalization. Moreover, *sir-2.4* RNAi only slightly impaired dauer formation, a process antagonized by IIS (Figure 4C). We conclude that that the effects of SIR-2.4 on DAF-16 are largely independent of IIS, and are most functionally significant in the context of stress.

SIR-2.4 regulates DAF-16 acetylation and localization independently of its catalytic function

In mammals, SIRT1 and other sirtuins deacetylate FoxO proteins to promote stress responses and other processes [4,6,30,31,32,33,36]. Although DAF-16 interacts with the acetyltransferases CBP and p300 [46], acetylation of DAF-16 itself has never been demonstrated. Given our data linking SIR-2.4 to stress resistance and DAF-16 localization and function, we tested whether DAF-16 was acetylated, and whether SIR-2.4 might play a role in regulating levels of this modification. Indeed, we were able to detect DAF-16 acetylation in *C. elegans*, and levels of this modification were elevated in *sir-2.4* KO (Figure 5A) or KD worms (Figure S4A). Moreover, reciprocal co-immunoprecipitation studies revealed that SIR-2.4 binds to DAF-16 in mammalian cells (Figure 5B), suggesting that SIR-2.4 might interact with DAF-16 to deacetylate this protein. To test this hypothesis, we performed *in vitro* deacetylation assay with purified SIR-2.4 proteins and pre-acetylated DAF-16 as substrates. However, in multiple experiments we were unable to detect significant deacetylation of DAF-16 by SIR-2.4 *in vitro*, despite efficiently deacetylating histones with mammalian SIRT1 and SIRT6 in parallel under identical conditions (data not shown).

Given this result, the requirement for SIR-2.4 catalytic function in DAF-16 localization was directly tested in worms. As shown above, *sir-2.4* KO worms showed impaired DAF-16 nuclear localization in response to stress. This defect was rescued by transgenic expression of wild-type SIR-2.4 (Figure 5C). Curiously, expression of the SIR-2.4 N124A mutant, bearing a mutation at highly conserved residues predicted based on homology to greatly diminish catalytic function [47], restored DAF-16 relocalization as efficiently as wild-type SIR-2.4. Unfortunately, co-overexpression of *DAF-16* together with chromosomally integrated *SIR-2.4* caused worms to be very sick and produce few progeny, precluding a direct assessment of DAF-16 acetylation in rescued animals (data not shown). We conclude that SIR-2.4 regulates DAF-16 acetylation and localization independently of its catalytic activity,

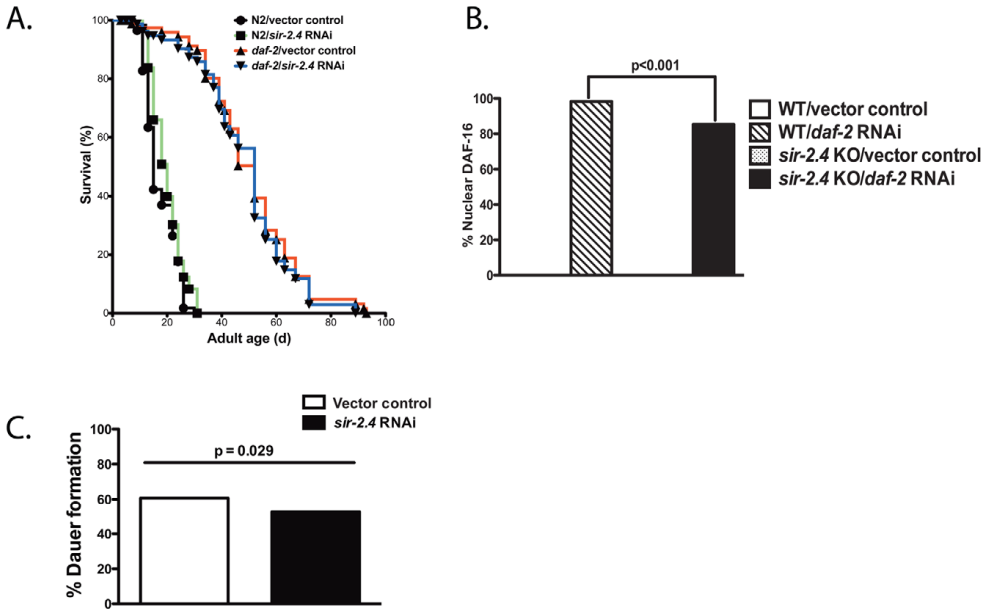


Figure 4. Minimal impact of SIR-2.4 on IIS-induced longevity, DAF-16 nuclear localization induced by reduced IIS, and dauer formation. (A) Lifespan analysis of wild-type (N2) animals or *daf-2(e1370)* mutants grown on vector control bacteria (black or red) or *sir-2.4* RNAi bacteria (green or blue) at 20°C. (B) DAF-16 nuclear localization was assessed in TJ356 (expressing *daf-16::gfp* in WT background) or EQ200 (expressing *daf-16::gfp* in *sir-2.4(n5137)* background) animals. Animals were fed with either vector control or *daf-2* RNAi from the time of hatching. Worms were scored for the presence or absence of GFP accumulation within the head hypodermic nuclei as day 1 adult (n = 116 or greater) under unstressed condition. *P*-values were calculated by Pearson's chi-square test. (C) *daf-2(e1370)* mutants (P_0) were fed with control or *sir-2.4* RNAi bacteria at 20°C. F_1 eggs were then moved to 22°C for 72 hours prior to being scored for dauer formation (n = 336 or greater). *P*-values were calculated by Pearson's chi-square test.
doi:10.1371/journal.pgen.1002948.g004

potentially by preventing the action of an acetyltransferase on DAF-16 and/or recruiting another deacetylase.

Potential cooperativity between SIR-2.1 and SIR-2.4 with respect to regulation of DAF-16 acetylation was assessed (Figure S4B). As before, *sir-2.4* KO caused a large increase in DAF-16 acetylation. Interestingly *sir-2.1* RNAi caused a modest but detectable increase in DAF-16 acetylation. In *sir-2.1* KD;*sir-2.4* KO animals, levels of DAF-16 acetylation resembled those in the *sir-2.4* KO. Thus, SIR-2.1 does impact DAF-16 acetylation, perhaps by directly deacetylating DAF-16, as has been shown for its homolog mammalian SIRT1. However, in contrast to SIR-2.4, the impact of SIR-2.1 on DAF-16 acetylation is quantitatively insufficient to promote significant DAF-16 nuclear localization (our results and [17]). Alternatively, it is possible that SIR-2.1 and SIR-2.4 modulate acetylation on different DAF-16 lysine residue that regulate divergent aspects of DAF-16 function.

SIR-2.4 blocks CBP1-dependent DAF-16 acetylation

Acetylation regulates of FoxO proteins in a context-specific manner [3]. To identify the enzyme that acetylates DAF-16 *in vivo*, five *C. elegans* acetyltransferases *CBP-1* (p300/CBP), *PCAF-1* (p300/CBP-associated factor), *MYS-1*, *MYS-2* and *MYS-3* (MYST histone acetyltransferases) were examined for potential impacts on nuclear translocation and acetylation of DAF-16. RNAi KD of *cbp-1* induced dramatic, virtually complete DAF-16 nuclear translocation under basal, unstressed conditions (Figure 6A). In contrast, inhibition of *pcaf*-1 or simultaneous KD of all three MYST acetyltransferases had no effect on DAF-16 localization (data not shown). This effect of *cbp-1* RNAi on DAF-16 nuclear localization was largely epistatic to the impact of *sir-2.4* KO (Figure 6A). These results are consistent with a recent report showing that mammalian CBP promotes cytoplasmic localization of FoxO3A [48]. Consistent with a role for CBP-1 in acetylating DAF-16 *in vivo*, we found that DAF-16 is hypo-acetylated in *cbp-1* KD animals (Figure 6B). To identify CBP-1 target sites on DAF-16, we performed mass spectrometry analysis on recombinant DAF-16 treated with CBP-1 *in vitro* (Figure 6C). We identified four CBP-1 dependent acetylation sites on DAF-16: K248, K253, K375 and K379 (Table S3). Together, these findings suggest that DAF-16 is acetylated by CBP-1 *in vivo* to promote its nuclear exclusion under basal conditions.

It has been reported that *cbp-1* RNAi confers stress sensitivity and reduced lifespan [49]. We confirmed that *cbp-1* RNAi caused hypersensitivity to heat shock (Figure S5A) and oxidative insult (Figure S5B). Moreover, as previously reported, *cbp-1* RNAi caused worms to be very short-lived (Figure S5C). We conclude that the constitutive nuclear localization of DAF-16 occurring in the context of *cbp-1* inhibition is inadequate to promote stress resistance or increased longevity, perhaps due to loss of *cbp-1* acetylation of other key targets necessary for health and normal

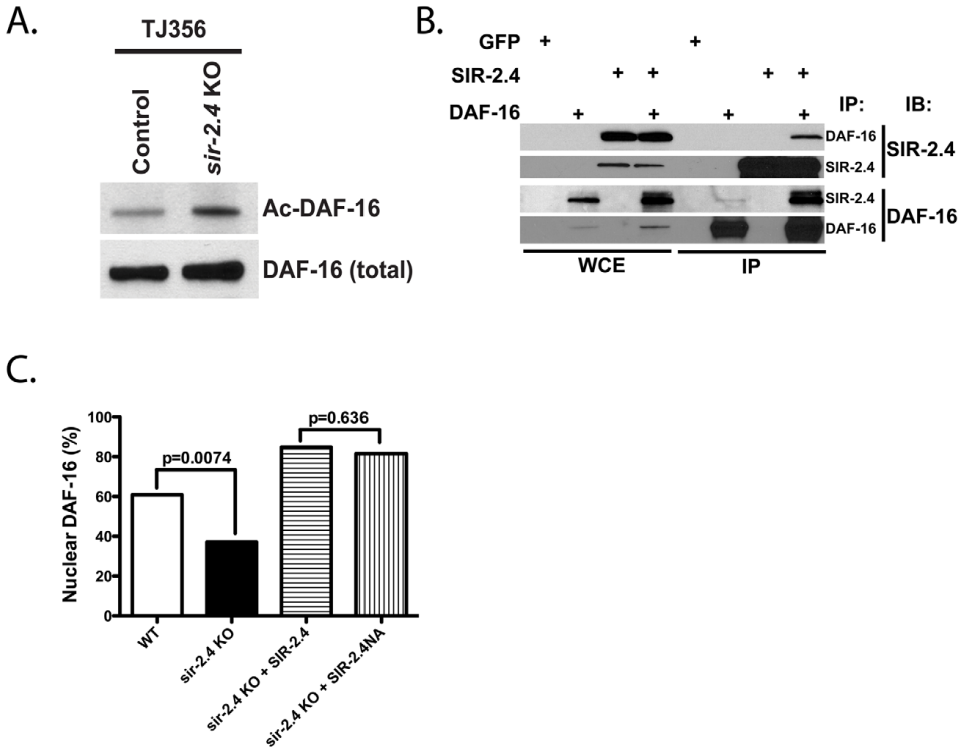


Figure 5. SIR-2.4 interacts with DAF-16 and promotes DAF-16 deacetylation and function independently of catalytic activity. (A) *sir-2.4* deletion promotes DAF-16 hyperacetylation. DAF-16 acetylation was assessed in control or *sir-2.4* KO worms by acetyl-lysine immunoprecipitation followed by GFP immunoblot. (B) SIR-2.4 and DAF-16 interact. Plasmids encoding FLAG-tagged SIR-2.4 and HA-tagged DAF-16 were transfected into 293T cells as indicated (GFP, negative control). Immunoprecipitation and immunoblotting were performed as shown. (C) Rescue of DAF-16 nuclear localization with a catalysis-defective *sir-2.4* mutant. Stable transgenic strains of *sir-2.4(n5137)* were generated expressing either wild-type SIR-2.4 or the *sir-2.4* N124A mutant. Worms were scored for GFP accumulation within the head hypodermic nuclei as day 1 adult (n = 50 or greater) after 20 min of heat-shock at 35°C. *P*-values were calculated by Pearson's chi-square test.
doi:10.1371/journal.pgen.1002948.g005

lifespan, potentially in many different cell types. Consistent with this hypothesis, CBP-1 is widely express in most or all somatic cells of the worms [50,51,52].

The DEK protein binds histones to prevent p300- and PCAF-mediated histone acetylation [53]. To test whether SIR-2.4 might exert its effect on DAF-16 through a similar mechanism, we examined the effect of the presence of SIR-2.4 on CBP-mediated DAF-16 acetylation *in vitro*. Purified DAF-16 and CBP were incubated with or without SIR-2.4. The presence of SIR-2.4 blocked CBP-mediated acetylation of DAF-16 (Figure 6C). A similar result was also obtained when DAF-16 and CBP were incubated with the SIR-2.4 N124A catalytic mutant (Figure 6C). We conclude that SIR-2.4 inhibits CBP-mediated DAF-16 acetylation likely through protein-protein interaction, independent of deacetylase function (Figure 7).

Discussion

We have shown that the sirtuin SIR-2.4 plays a crucial role in promoting DAF-16 activity and stress resistance. In response to multiple forms of stress, SIR-2.4 promotes DAF-16 nuclear recruitment and function, and ultimately organismal survival. SIR-2.4 is also required for optimal DAF-16 deacetylation and transcriptional activity under basal, unstressed conditions. In contrast, SIR-2.4 does not greatly impact DAF-16 nuclear localization, dauer formation, or longevity due to reduced IIS. We find that SIR-2.4 associates with DAF-16 and negatively regulates DAF-16 acetylation by interfering with the ability of CBP-1 to acetylate DAF-16 under stress conditions. Although mammalian FoxO proteins are acetylated in a functionally significant manner [3], to our knowledge this is the first demonstration that this modification is conserved in *C. elegans*. These data suggest that deacetylation likely plays two distinct roles in modulating DAF-16 function, in promoting optimal DAF-16 transcriptional activity under both basal and stress conditions, and in promoting DAF-16 nuclear recruitment following stress. We propose a model in which SIR-2.4 does not interact with DAF-16 under unstressed conditions. This allows CBP-1 to acetylate DAF-16 and sequester DAF-16 in the cytosol. In response to stress, SIR-2.4 binds to DAF-16 to block CBP-1-depedent acetylation and

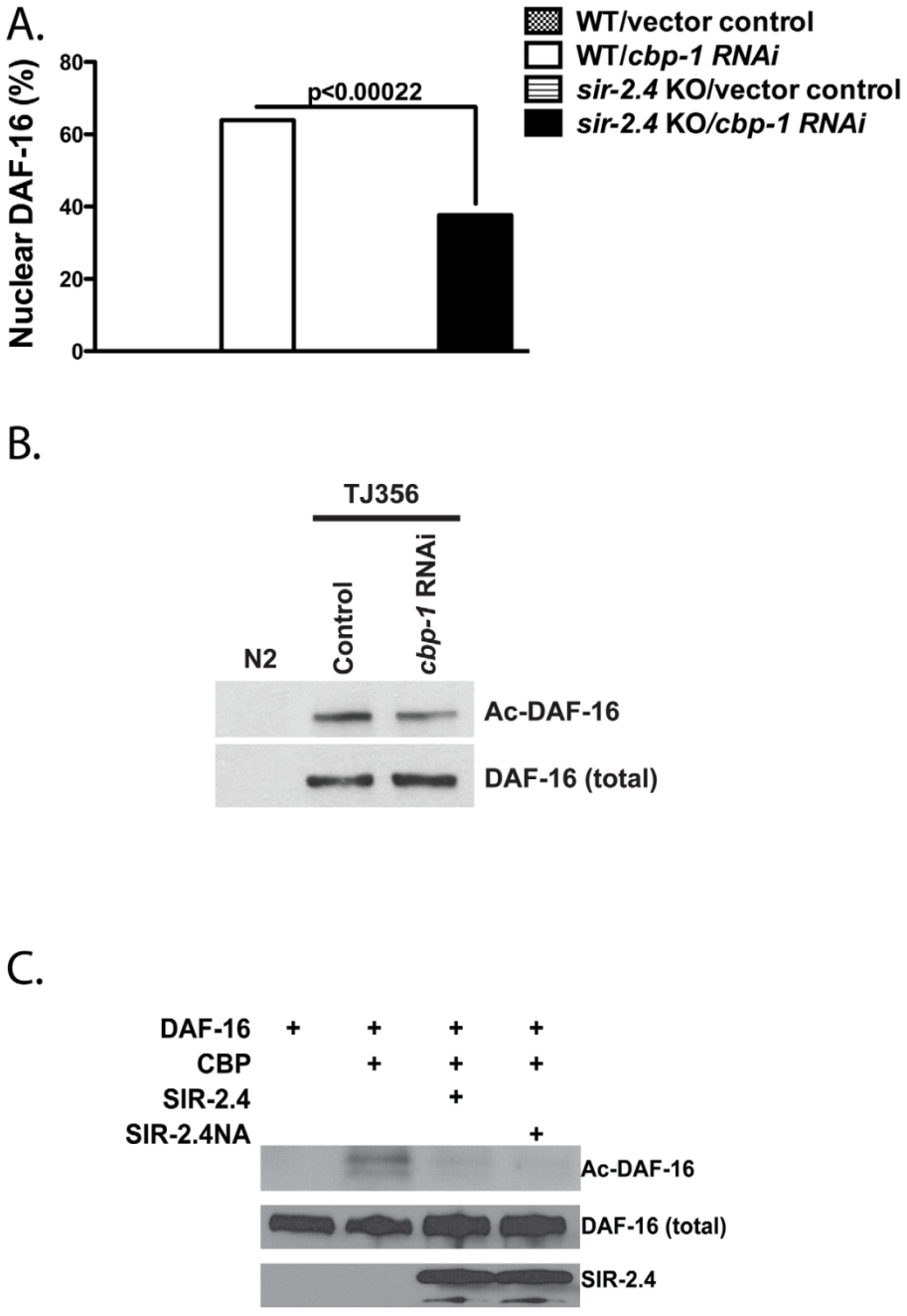

A.
WT/vector control
WT/cbp-1 RNAi
sir-2.4 KO/vector control
sir-2.4 KO/cbp-1 RNAi
Nuclear DAF-16 (%)
80
60
40
20
0
p<0.00022
B.
TJ356
N2
Control
cbp-1 RNAi
Ac-DAF-16
DAF-16 (total)
C.
DAF-16 + + + +
CBP + + +
SIR-2.4 +
SIR-2.4NA +
Ac-DAF-16
DAF-16 (total)
SIR-2.4

Figure 6. SIR-2.4 inhibits CBP1-mediated DAF-16 acetylation. (A) DAF-16 nuclear localization was assessed in TJ356 (expressing *daf-16::gfp* in WT background) or EQ200 (expressing *daf-16::gfp* in *sir-2.4(n5137)* background) animals. Animals (n = 90 or greater) were scored for DAF-16::GFP nuclear translocation as described in Figure 4B. (B) *cbp-1* KD decreases DAF-16 acetylation. DAF-16 acetylation was assessed in TJ356 worms grown on either control or *cbp-1* RNAi by acetyl-lysine immunoprecipitation followed by GFP immunoblot. (C) SIR-2.4 blocks CBP1-dependent DAF-16 acetylation *in vitro*. Purified DAF-16 was incubated with CBP in the presence of WT SIR-2.4 or the SIR-2.4 NA mutant at 37°C. DAF-16 acetylation levels were assessed as described in (B).
doi:10.1371/journal.pgen.1002948.g006

promote DAF-16 nuclear translocation and gene expression (Figure 7).

There are several important issues raised by our results. First, our data suggest that deacetylation promotes DAF-16 nuclear localization and function in response to stress, but is largely dispensable in the context of reduced IIS, suggesting that regulation of this modification may be involved in adjusting DAF-16 activity in the face of different organismal requirements. Our mass spectrometry analysis has identified four acetylation sites on DAF-16 (Table S3). Notably, acetylation at one of these (K248), is also present at homologous lysine residues in FoxO1 (K262) and FoxO3A (K259) [4,54,55,56]. Future studies mutating these acetylation sites individually and in combinations will be required to identify specific residues involved in regulating DAF-16 nuclear recruitment and transcriptional activity.

Our results provide further evidence that multiple sirtuins regulate FoxO proteins by diverse mechanisms to achieve different functional outcomes. In mammals, several sirtuins directly deacetylate FoxO proteins [4,6,30,31,32,33,36]. In worms, SIR-2.1 promotes DAF-16 function to increase stress resistance via 14-3-3 proteins [26]. We find SIR-2.4 suppresses DAF-16 acetylation and is critical in the organismal response to a variety of stressors. Whereas SIR-2.4 is required for optimal DAF-16 nuclear recruitment in response to stress, SIR-2.1 is dispensable for this process (Figure 1C-1D) [17]. It has been reported that SIRT1 promotes survival of cerebellar granular neurons in response to low extracellular potassium levels in a catalytic activity-independent fashion [57]. Thus, sirtuins may regulate stress responses through mechanisms other than direct covalent substrate modification. Consistent with this idea, we found that SIR-2.4 regulates DAF-16 function independent of its catalytic function, by preventing CBP-1-mediated DAF-16 acetylation. We note that SIRT6 was recently shown to be capable of binding NAD^+ and undergoing a conformational change in the absence of substrate; based on this finding, it has been speculated that SIRT6 might act as an NAD^+ metabolite sensor, independently of catalytic activity [58]. This hypothesis is consistent with our data regarding the SIRT6 homolog SIR-2.4. Chromatin modifying complexes frequently involve multiple enzymes with apparently redundant biochemical activities, including multiple histone deacetylases; analogously SIR-2.4 likely associate with other deacetylases to regulate DAF-16. SIR-2.1 associates with DAF-16, although SIR-2.1 was not shown to deacetylate DAF-16 [21,26]. DAF-16 deacetylases could in principle be SIR-2.1 or other worm sirtuins, or non-sirtuin deacetylases. However, the fact that SIR-2.1 does not play a major role in stress-induced DAF-16 nuclear recruitment strongly suggests that SIR-2.1 is not likely the major DAF-16 deacetylase.

FoXO proteins has been shown to interact with acetyltransferases CBP and p300 in mammalian cells [46], although it was previously unknown whether these proteins directly acetylate DAF-16 in worms. In this study, we showed that inactivation of *cbp-1* by RNAi results in constitutive nuclear localization of DAF-16, and that CBP-1 can directly acetylate DAF-16 on multiple residues *in vitro* (Figure 6C, Table S3). However, *cbp-1* KD only modestly reduces the level of DAF-16 acetylation *in vivo* (Figure 6B). Since the pan-acetyl-lysine antibody we used in these experiments is not able to discriminate among DAF-16 species acetylated at different sites, or singly versus multiply acetylated DAF-16 species, it is possible that CBP-1 is not the only acetyltransferase that acetylates DAF-16 *in vivo*. Thus the effect of *cbp-1* KD on DAF-16 acetylation may be partially obscured by CBP-1-independent acetylation of DAF-16. Nevertheless, CBP-1 has a major impact on DAF-16 localization (Figure 6A).

Finally, it is currently unclear whether SIR-2.4 regulates DAF-16 in a cell-autonomous matter, or whether SIR-2.4 is required in only a subset of cells to carry out this function. The tissue expression pattern of endogenous SIR-2.4 is ambiguous. A SIR-2.4::GFP translational fusion protein shows weak expression in most cell types, but much stronger expression in a subset of head and tail neurons as well as in a subset of somatic gonad cells (Figure S1). Thus, SIR-2.4 could conceivably regulate DAF-16 in

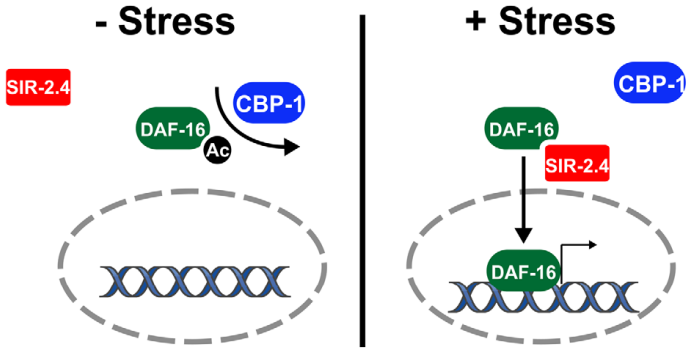


Figure 7. Model: SIR-2.4 promotes DAF-16 deacetylation and function during stress. See text for details.
doi:10.1371/journal.pgen.1002948.g007

neurons to influence the systemic hormonal milieu *e.g.*, by reducing INS-7 expression thereby impacting DAF-16 localization and function throughout the animal. However, our observation that DAF-16 is hyperacetylated in extracts from whole worms, together with our biochemical data, supports the hypothesis that SIR-2.4 regulates DAF-16 acetylation in most tissues of the animal cell autonomously. Overall, SIR-2.4 is a novel regulator of DAF-16 localization and function in response to stress in *C. elegans*.

Materials and Methods

Strains

CF1041: *daf-2(e1370)*III, TJ356: *zIs356* [*daf-16::gfp* + *rol-6*], AM140: *rmIs132*[*unc-54p::Q35::yfp*], MT18068: *sir-2.4(n5137)*I, EQ 137: *iqEx*47 [*sir-2.4p::sir-2.4::gfp* + *rol-6*], EQ158: *iqEx*50 [*sir-2.4p::sir-2.4* + *myo-3p::rfp*], EQ 200: *sir-2.4(n5137)*I; *zIs356* [*daf-16::gfp* + *rol-6*], EQ205: *sir-2.4(n5137)*I; *zIs356* [*daf-16::gfp* + *rol-6*]; iqEx59 [*sir-2.4p::sir-2.4NA* + *myo-3p::rfp*], EQ211: *sir-2.4(n5137)*I; *zIs356* [*daf-16::gfp* + *rol-6*]; iqEx60 [*sir-2.4p::sir-2.4* + *myo-3p::rfp*].

All strains used were maintained and handled as described previously [59]. TJ356, AM140 and CF1041 were obtained from the Caenorhabditis Genetic Center. For the generation of transgenic animals, plasmid DNA mixes were injected into the gonad of young adult hermaphrodite animals, using the standard method described previously [60]. F_1 progeny were selected on the basis of the roller phenotype. Individual F_2 progenies were isolated to establish independent lines. For the generation of the EQ137 (SIR-2.4::GFP overexpressor) strain, plasmid DNA containing a mixture of 100 ng/ml of *sir-2.4p::sir-2.4::gfp* and 50 ng/ml of pRF4 (*rol-6*) constructs was injected into N2 animals. For the generation of EQ158 (native SIR-2.4 overexpressor), plasmid DNA containing a mixture of 30 ng/ml of native *SIR-2.4* driven by its own promoter and 50 ng/ml of coinjection marker *myo-3p::rfp* was injected into N2 animals. For the generation of EQ211, plasmid DNA containing a mixture of 30 ng/ml of *sir-2.4p::sir-2.4* and 80 ng/ml of *myo-3p::rfp* constructs was injected into TJ356 animals. For the generation of EQ205, plasmid DNA containing a mixture of 30 ng/ml of *sir-2.4p::sir-2.4NA* and 80 ng/ml of *myo-3p::rfp* constructs was injected into TJ356 animals.

RNA–interference (RNAi) experiments

HT115 bacteria transformed with RNAi vectors (L4440) expressing dsRNA of the genes indicated were grown at 37°C in LB with 10 mg/mL tetracycline and 50 mg/mL carbenicillin, then seeded onto NG-carbenicillin plates and supplemented with 100 µL 0.1 M IPTG. The *sir-2.4* RNAi construct was generated by cloning nucleotides 1 467 of the *SIR-2.4* cDNA into the L4440 vector. The identity of all RNAi clones was verified by sequencing the inserts using M13-forward primer. Eggs were added to plates and transferred to new plates every 3 6 days.

sir-2.4 deletion mutant analysis

The deletion in a *sir-2.4* mutant strain (kind gift of the Horvitz laboratory) was mapped by PCR, and found to encompass 1,929 bp of chromosome I (5990775 5992703), which encodes nucleotides 28 879 of the SIR-2.4 spliced mRNA (data available upon request).

DAF-16 nuclear localization assay

For quantification of DAF-16::GFP localization, synchronized eggs from TJ356 animals (*i.e.* transgenic animals expressing DAF-16::GFP) or other strains as indicated were seeded onto either vector control or appropriate RNAi plates. For stress response experiments, day 1 adults were washed with M9 three times and transferred to new plates or and subjected to heat shock (35°C) or oxidative stress (1.5 mM H_2O_2 in M9). GFP localization was then analyzed using an Olympus BX61 fluorescent microscope at 40× or 100× magnification. For time-course analysis, worms were scored for the presence or absence of GFP accumulation within the nuclei of head hypodermis cells (n = 30∼50) in a blinded fashion every 5 30 min. An animal was scored as having nuclear GFP if more than one head hypodermic nuclei contained DAF-16::GFP. For single time point experiments, worms were blindly scored for the presence or absence of GFP accumulation within the nuclei of indicated cells (n = 120 or greater). *P* values were calculated by Poisson regression (time-course assays) or chi-square test (single time point assays).

Quantitative RT–PCR analysis

Total RNA was isolated from approximately 5,000 day 1 adult worms, and cDNA was generated from 4 µg of RNA using Superscript III RT (Invitrogen). Real-time qRT-PCR experiments were performed using the Power SYBR Green PCR Master mix (Applied Biosystems) and the Chromo 4 system (MJ Research). Relative mRNA level of the genes of interest were calculated and normalized against an internal control (ACT-1; worm β-actin). Primer sequences were (all 5′-3′): SOD-3 (GTTTCAGCGC-GACTTCGGTTCCCT, CGTGCTCCCAAACGTCAATTC-CAA); DOD-3 (AAAAAGCCATGTTCCCGAAT, GCTGCGA AAAGCAAGAAAAT); DOD-24 (TGTCCAACACAACCTG-CATT, TGTGTCCCGAGTAACAACCA); C32H11.4 (TTACT TCCCATCGCCAAAGT, CAATTCCGGCGATGTATGAT); HSP-16.1 (GATCAAAAGTTTGCCATAAATCTC, TTCAGT CTTTAATTCTTGTTCTCC); INS-7 (TCGTTGTGGAAGAA GAATACATTC, TTAAGGACAGCACTGTTTTCG); and AC T-1 (CTACGAACTTCCTGACGGACAAG, CCGGCGGACT CCATACC).

Stress assays

For thermotolerance assays, eggs from N2 worms were transferred to plates seeded with vector control, *daf-16* RNAi, or *sir-2.4* RNAi bacteria and grown to day 1 of adulthood. Worms were then transferred to plates without any food and heat-shocked at 35°C. Viability was determined at the indicated timepoints; death was determined by the lack of movement after prodding. For oxidative stress assays, eggs from N2 worms were transferred to plates seeded with vector control, *daf-16* RNAi, or *sir-2.4* RNAi bacteria and grown to day 3 of adulthood. Worms were then transferred to 24-well plates and soaked in 1.5 mM H_2O_2 in M9 media. Viability was determined at the indicated timepoints as above. For stress assays, a Kaplan Meier survival analysis with a log-rank test was performed, and a *P*-value of 0.05 was considered statistically significant.

Paralysis assay

Synchronized eggs from AM140 animals (i.e. transgenic animals expressing Q35::YFP) were seeded on either vector control or the indicated RNAi bacteria. Animals were scored for polyQ-induced paralysis every other day during adulthood. Paralyzed worms were identified as those failing to make forward or backward movement in response to stimulation by plate-tapping and tail-prodding; these worms still exhibited pharyngeal pumping. For the paralysis assay, a Kaplan Meier survival analysis with a log-rank test was performed.

Lifespan analysis

Lifespan analysis were conducted at 20°C as described previously [61,62]. Strains were grown at 20°C for at least two

generations without starvation prior to lifespan analysis. At least 60 worms were used for each experiment. In all experiments, the pre-fertile period of adulthood was used as t = 0 for lifespan analysis. Stata 8 software was used for statistical analysis to determine the means and percentiles. In all cases, *P*-values were calculated using the log-rank (Mantel-Cox) method.

Assessment of DAF-16 acetylation in worms

~15,000 synchronized day 1 adult worms grown at 20°C were harvested by washing three times with cold M9 buffer and once with HB-high salt buffer (10 mM HEPES, pH 7.9; 10 mM KCl; 1.5 mM $MgCl_2$; 0.1 mM EDTA; 0.5 mM EGTA; 44 mM Sucrose; 100 mM NaCl; 0.5% Triton X-100). Worm pellets were then resuspended in 3× volume of HB-high salt buffer supplemented with Protease Inhibitor Cocktail Complete (Roche), 20 mM β-glycerophosphate, 1 mM sodium orthovanadate, 1 mM nicotinamide and 1 µM trichostatin A. Pellets were immediately frozen and stored in liquid nitrogen. Frozen suspensions were thawed, homogenized with a Dounce homogenizer (30 strokes with pestle B), and centrifuged at 14,000×g at 4°C for 20 minutes. Supernatants were collected and total protein concentrations were quantified by Bradford assay. For immunoprecipitation, 30 µl of anti-acetyl-lysine agarose beads (Immunechem; ICP0388) were added to 1 mg of protein extract with 100 µg/ml of ethidium bromide and incubated with gentle shaking at 4°C overnight. The beads were then washed 3 times with HB-high salt buffer supplemented with 50 mg/ml ABESF, 1 mM sodium orthovanadate, 1 mM nicotinamide and 1 µM trichostatin A. before being subjected to western blotting analysis. The samples were subjected to SDS-PAGE and transferred to a PVDF membrane (Millipore). The membrane was washed three times with TBS containing 0.1% Tween 20 (TBST). After blocking with TBST containing 5% nonfat milk for 60 min, the membrane was incubated with the primary antibody indicated (e.g. anti-GFP, Abcam, #AB6556) at 4°C for 12 h and washed three times with TBST. The membrane was then probed with HRP-conjugated secondary antibody for 1 h at room temperature and washed with TBST three times. Finally, the immunoblots were developed using a chemiluminescent substrate (Millipore) and visualized by autoradiography.

Assessment of SIR-2.4/DAF-16 interaction

10 cm dishes of HEK293T cells were transfected using TransIT-293 (Mirus) with plasmids encoding HA-tagged DAF-16 (1 µg), FLAG-tagged SIR-2.4 (9 µg), both DAF-16 and SIR-2.4 plasmids together, or a plasmid encoding GFP (5 µg), to assess transfection efficiency. Cells were harvested 48 hours post transfection, and were lysed by rotation at 4°C for 20 minutes in lysis buffer (LB; 150 mM NaCl, 1% Triton X-100, 0.5% NP40, 50 mM Tris pH 7.4, 10% glycerol, with Protease Inhibitor Cocktail Complete-EDTA free (Roche) [63], followed by brief sonication. 1 mg of whole cell extract from each cell line was pre-cleared by slow rotation with 50 µl of protein G conjugated agarose beads. For immunoprecipitation, either 25 µl of M2-agarose beads (Sigma) (anti-FLAG IP) or 25 µl of anti-HA-agarose beads (Roche) (anti-HA IP) was added to pre-cleared WCE, and IPs were incubated overnight by slow rotation at 4°C. After incubation, beads were pelleted and washed three times in LB. FLAG-tagged proteins were eluted in 80 µl of FLAG elution buffer (150 ng/µl 3×FLAG peptide (Sigma), 10 mM Tris-HCl (pH 8.0) and 150 mM NaCl) at 4°C for 4 hours. 40 µl of eluate was loaded on an SDS-PAGE gel. For HA IPs, 80 µl of Laemmli buffer was added to the anti-HA-agarose beads, and beads were boiled at 100°C for 5 minutes; 40 µl was loaded on a gel. DAF16 and SIR2.4 interaction was assessed by immunoblot using anti-DAF-16 (Santa Cruz) or anti-HA antibodies.

In vitro DAF-16 acetylation assay

3×FLAG tagged DAF-16, CBP or SIR-2.4 were each transfected and expressed in 293T cells. These proteins were then purified with Anti-FLAG M2 Affinity Gel (Sigma, A2220). 0.5 µg of purified DAF-16 was incubated with 0.2 µg of purified CBP with or without 0.5 µg of purified SIR-2.4 in the presence of HAT buffer (50 mM Tris-HCl, pH 8.0, 0.1 mM EDTA, 1 mM dithiothreitol, 10%glycerol) supplemented with 200 µM of acetyl-CoA. The reactions were incubated at 37°C for 3 hours and analyzed by western blot. Acetylated DAF-16 was queried with anti-acetyl-lysine antibody (Immunechem; ICP0380) and total level of DAF-16 were assessed with anti-FLAG antibody (Sigma, #F3165).

Supporting Information

Figure S1 Expression pattern of *sir-2.4* in *C. elegans*. Transgenic lines expressing a SIR-2.4 translational GFP fusion (*sir-2.4p::sir-2.4::gfp*) were utilized to analyze the expression pattern of *sir-2.4* in *C. elegans*. Images of (a b) L3 or (c e) late L4/young adult stage transgenic animals (EQ137) expressing GFP protein under control of the *sir-2.4* promoter. *sir-2.4* is highly expressed in a subset of head and tail neurons beginning at early larval stage, indicated by white arrows. High expression of *sir-2.4* is also found in spermathecal-uterine valve (sp-ut valve) cells beginning at L4 larval stage, indicated by white arrows. The yellow arrow indicates the nuclear accumulation of SIR-2.4::GFP fusions in these cells. It is worth noting that very weak expression of *sir-2.4* is found ubiquitously in most tissues, although it is difficult to capture in these images.
(PDF)

Figure S2 Effects of *sir-2.4* RNAi or deletion on gene expression, stress resistance, and IIS-mediated DAF-16 translocation. (A B) Expression of the genes indicated was measured in worms of the indicated genotypes as in Figure 2. (C) Thermotolerance was assessed in worms of the indicated genotypes as in Figure 3A. (D) Oxidative stress resistance was assessed in worms of the indicated genotypes as in Figure 3B. (E) TJ356 animals were fed with either vector control, *daf-2* RNAi, a 1:1 mix of *daf-2* and *sir-2.4* RNAi, or a 1:1 mix of *daf-2* and *yk615e* RNAi bacteria from the time of hatching. *yk615e* is a gene randomly selected as a negative control in double RNAi experiments. Animals (n = 125 or greater) were scored for DAF-16::GFP nuclear translocation as described in Figure 4B.
(PDF)

Figure S3 Minimal impact of SIR-2.4 overexpression on stress resistance or lifespan. Transgenic animals overexpressing (A) a SIR-2.4::GFP fusion protein or (B) native untagged SIR-2.4 protein and N2 controls were exposed to 35°C heat stress. Viability was then scored at the timepoints indicated. (C) Survival curves of wild-type (N2) animals or transgenic animals overexpressing native SIR-2.4 (EQ158) at 20°C. All statistical details are presented in Tables S1 and S2.
(PDF)

Figure S4 *sir-2.4* RNAi promotes DAF-16 hyperacetylation. (A) DAF-16 acetylation was assessed in control or *sir-2.4* RNAi worms by acetyl-lysine immunoprecipitation followed by GFP immunoblot as described in Figure 5A. (B) DAF-16 acetylation was assessed in *sir-2.4* KO, *sir-2.1* RNAi, or double loss-of function animals as indicated.
(PDF)

Figure S5 *cbp-1* loss of function confers stress sensitivity and shortened lifespan. Mean survival of *cbp-1* RNAi worms in response to heat shock (A) or peroxide stress (B). (C) Lifespan curves of wildtype or *sir-2.4* KO animals in the presence of *cbp-1* or control RNAi bacteria.
(PDF)

Table S1 Effects of sir-2.4 expression on stress resistance. Animals indicated were exposed to heat or oxidative stress. Mean survival ± SEM, in hours, observed in the stress analysis is shown. 75th percentile is the time at which the fraction of animals alive reaches 0.25. 'n' indicates the number of animals scored in the each experiment. *P*-Values calculated by pair-wise comparisons to vector control of the same experiment. We used Stata 8 software for statistical analysis and to determine means and percentiles. The logrank (Mantel-Cox) test was used to test the hypothesis that the survival functions among groups were equal. a *P*-Values calculated by pair-wise comparisons to N2 grown on vector control of the same experiment. b Compared to N2 grown on the same RNAi bacteria. c Compared to the same mutants grown on vector control. '*' indicates the sets of experiments plotted and shown in Figures.
(PDF)

Table S2 Effects of *SIR-2.4* on lifespan. Adult mean lifespan ± SEM, in days, observed in lifespan analyses. Lifespan experiments were carried out at 20°C. 75th percentile is the age at which the fraction of animals alive reaches 0.25. 'n' shows the number of observed deaths relative to total number of animals started at day 1. The difference between these numbers represents the number of animals censored during the experiment, and includes animals that exploded, bagged (i.e. exhibited internal progeny hatching), or crawled off the plates. a *p*-Values calculated by pair-wise comparisons to N2 control of the same experiment. b p-Values calculated by pair-wise comparisons to mutants fed with control bacteria of the same experiment. We used Stata 8 software for statistical analysis and to determine means and percentiles. The logrank (Mantel-Cox) test was used to test the hypothesis that the survival functions among groups were equal.
(PDF)

Table S3 Peptides representing unique acetylated sequences from the FOXO transcription factor are shown with additional information. A"#" sign indicates the site of acetylation in the sequence. The site position refers to which lysine residue on FOXO the identified acetylation site represents. The mass to charge ratio (m/z) and charge state of the peptide as they were observed in the mass spectrometer are given. The mass error specifies the difference between the observed m/z ratio and the theoretical m/z ratio of a peptide, reported in part per million (PPM). The XCorr is the cross correlation score between the theoretical and observed MS^2 spectra for the matched peptide. The unique ΔCorr is the difference between the XCorr of the top ranking peptide match (the reported peptide) and the XCorr of the next closest ranked peptide of a unique primary sequence (not simply an alternate placement of the acetylation site), normalized by the XCorr of the top ranking peptide match. All reported peptides passed a cutoff of a false discovery rate (FDR) <0.1%, based on the target-decoy strategy [61]. See Text S1 for materials and methods used in these studies.
(DOCX)

Text S1 Supplemental materials and methods.
(DOCX)

Acknowledgments

The authors thank H. R. Horvitz for providing the MT18068 strain; the Ruvkun lab for providing the *daf-16* RNAi plasmid; the *Caenorhabditis* Genetic Center for the TJ356, AM140, and CF1041 strains; and the National Biosource Project of Japan (Dr. Mitani) for the *sir-2.4 (tm2817)* strain. Scott Pletcher is gratefully acknowledged for helpful discussions and comments on the manuscript. Hsiu-yin Chiang is gratefully acknowledged for statistical analysis. The authors also thank Carol Mousigan for technical assistance.

Author Contributions

Conceived and designed the experiments: A-LH DBL. Performed the experiments: W-CC DXT BY JW-G XY TM ME A-LH. Analyzed the data: W-CC DXT BY JW-G XY SPG DBL A-LH . Wrote the paper: DBL A-LH.

References

1. Arden KC (2008) FOXO animal models reveal a variety of diverse roles for FOXO transcription factors. Oncogene 27: 2345 2350.
2. Yen K, Narasimhan SD, Tissenbaum HA (2011) DAF-16/Forkhead box O transcription factor: many paths to a single Fork(head) in the road. Antioxid Redox Signal 14: 623 634.
3. Calnan DR, Brunet A (2008) The FoxO code. Oncogene 27: 2276 2288.
4. Brunet A, Sweeney LB, Sturgill JF, Chua KF, Greer PL, et al. (2004) Stress-dependent regulation of FOXO transcription factors by the SIRT1 deacetylase. Science 303: 2011 2015.
5. Frescas D, Valenti L, Accili D (2005) Nuclear trapping of the forkhead transcription factor FoxO1 via Sirt-dependent deacetylation promotes expression of glucogenetic genes. J Biol Chem 280: 20589 20595.
6. van der Horst A, Tertoolen LG, de Vries-Smits LM, Frye RA, Medema RH, et al. (2004) FOXO4 is acetylated upon peroxide stress and deacetylated by the longevity protein hSir2(SIRT1). J Biol Chem 279: 28873 28879.
7. Kitamura YI, Kitamura T, Kruse JP, Raum JC, Stein R, et al. (2005) FoxO1 protects against pancreatic beta cell failure through NeuroD and MafA induction. Cell Metab 2: 153 163.
8. van der Horst A, de Vries-Smits AM, Brenkman AB, van Triest MH, van den Broek N, et al. (2006) FOXO4 transcriptional activity is regulated by monoubiquitination and USP7/HAUSP. Nat Cell Biol 8: 1064 1073.
9. Lehtinen MK, Yuan Z, Boag PR, Yang Y, Villen J, et al. (2006) A conserved MST-FOXO signaling pathway mediates oxidative-stress responses and extends life span. Cell 125: 987 1001.
10. Essers MA, Weijzen S, de Vries-Smits AM, Saarloos I, de Ruiter ND, et al. (2004) FOXO transcription factor activation by oxidative stress mediated by the small GTPase Ral and JNK. EMBO J 23: 4802 4812.
11. Oh SW, Mukhopadhyay A, Svrzikapa N, Jiang F, Davis RJ, et al. (2005) JNK regulates lifespan in Caenorhabditis elegans by modulating nuclear translocation of forkhead transcription factor/DAF-16. Proc Natl Acad Sci U S A 102: 4494 4499.
12. Sunayama J, Tsuruta F, Masuyama N, Gotoh Y (2005) JNK antagonizes Akt-mediated survival signals by phosphorylating 14-3-3. J Cell Biol 170: 295 304.
13. Finkel T, Deng CX, Mostoslavsky R (2009) Recent progress in the biology and physiology of sirtuins. Nature 460: 587 591.
14. Spange S, Wagner T, Heinzel T, Kramer OH (2009) Acetylation of non-histone proteins modulates cellular signalling at multiple levels. Int J Biochem Cell Biol 41: 185 198.
15. Frye RA (2000) Phylogenetic classification of prokaryotic and eukaryotic Sir2-like proteins. Biochem Biophys Res Commun 273: 793 798.
16. Rizki G, Iwata TN, Li J, Riedel CG, Picard CL, et al. (2011) The evolutionarily conserved longevity determinants HCF-1 and SIR-2.1/SIRT1 collaborate to regulate DAF-16/FOXO. PLoS Genet 7: e1002235. doi:10.1371/journal.pgen.1002235
17. Heidler T, Hartwig K, Daniel H, Wenzel U (2010) Caenorhabditis elegans lifespan extension caused by treatment with an orally active ROS-generator is dependent on DAF-16 and SIR-2.1. Biogerontology 11: 183 195.
18. Pasco MY, Catoire H, Parker JA, Brais B, Rouleau GA, et al. (2010) Cross-talk between canonical Wnt signaling and the sirtuin-FoxO longevity pathway to protect against muscular pathology induced by mutant PABPN1 expression in C. elegans. Neurobiol Dis 38: 425 433.
19. Catoire H, Pasco MY, Abu-Baker A, Holbert S, Tourette C, et al. (2008) Sirtuin inhibition protects from the polyalanine muscular dystrophy protein PABPN1. Hum Mol Genet 17: 2108 2117.

20. Greiss S, Hall J, Ahmed S, Gartner A (2008) C. elegans SIR-2.1 translocation is linked to a proapoptotic pathway parallel to cep-1/p53 during DNA damage-induced apoptosis. Genes Dev 22: 2831 2842.
21. Wang Y, Oh SW, Deplancke B, Luo J, Walhout AJ, et al. (2006) C. elegans 14-3-3 proteins regulate life span and interact with SIR-2.1 and DAF-16/FOXO. Mech Ageing Dev 127: 741 747.
22. Viswanathan M, Kim SK, Berdichevsky A, Guarente L (2005) A role for SIR-2.1 regulation of ER stress response genes in determining C. elegans life span. Dev Cell 9: 605 615.
23. Parker JA, Arango M, Abderrahmane S, Lambert E, Tourette C, et al. (2005) Resveratrol rescues mutant polyglutamine cytotoxicity in nematode and mammalian neurons. Nat Genet 37: 349 350.
24. Mair W, Panowski SH, Shaw RJ, Dillin A (2009) Optimizing dietary restriction for genetic epistasis analysis and gene discovery in C. elegans. PLoS ONE 4: e4535. doi:10.1371/journal.pone.0004535
25. Tissenbaum HA, Guarente L (2001) Increased dosage of a sir-2 gene extends lifespan in Caenorhabditis elegans. Nature 410: 227 230.
26. Berdichevsky A, Viswanathan M, Horvitz HR, Guarente L (2006) C. elegans SIR-2.1 Interacts with 14-3-3 Proteins to Activate DAF-16 and Extend Life Span. Cell 125: 1165 1177.
27. Viswanathan M, Guarente L (2011) Regulation of Caenorhabditis elegans lifespan by sir-2.1 transgenes. Nature 477: E1 2.
28. Burnett C, Valentini S, Cabreiro F, Goss M, Somogyvari M, et al. (2011) Absence of effects of Sir2 overexpression on lifespan in C. elegans and Drosophila. Nature 477: 482 485.
29. Lombard DB, Pletcher SD, Canto C, Auwerx J (2011) Ageing: longevity hits a roadblock. Nature 477: 410 411.
30. Daitoku H, Hatta M, Matsuzaki H, Aratani S, Ohshima T, et al. (2004) Silent information regulator 2 potentiates Foxo1-mediated transcription through its deacetylase activity. Proc Natl Acad Sci U S A 101: 10042 10047.
31. Motta MC, Divecha N, Lemieux M, Kamel C, Chen D, et al. (2004) Mammalian SIRT1 represses forkhead transcription factors. Cell 116: 551 563.
32. Jing E, Gesta S, Kahn CR (2007) SIRT2 regulates adipocyte differentiation through FoxO1 acetylation/deacetylation. Cell Metab 6: 105 114.
33. Wang F, Tong Q (2009) SIRT2 suppresses adipocyte differentiation by deacetylating FOXO1 and enhancing FOXO1's repressive interaction with PPARgamma. Mol Biol Cell 20: 801 808.
34. Wang F, Nguyen M, Qin FX, Tong Q (2007) SIRT2 deacetylates FOXO3a in response to oxidative stress and caloric restriction. Aging Cell 6: 505 14.
35. Matsuzaki H, Daitoku H, Hatta M, Aoyama H, Yoshimochi K, et al. (2005) Acetylation of Foxo1 alters its DNA-binding ability and sensitivity to phosphorylation. Proc Natl Acad Sci U S A 102: 11278 11283.
36. Jacobs KM, Pennington JD, Bisht KS, Aykin-Burns N, Kim HS, et al. (2008) SIRT3 interacts with the daf-16 homolog FOXO3a in the mitochondria, as well as increases FOXO3a dependent gene expression. Int J Biol Sci 4: 291 299.
37. Sundaresan NR, Gupta M, Kim G, Rajamohan SB, Isbatan A, et al. (2009) Sirt3 blocks the cardiac hypertrophic response by augmenting Foxo3a-dependent antioxidant defense mechanisms in mice. J Clin Invest 119: 2758 2771.
38. Vakhrusheva O, Braeuer D, Liu Z, Braun T, Bober E (2008) Sirt7-dependent inhibition of cell growth and proliferation might be instrumental to mediate tissue integrity during aging. J Physiol Pharmacol 59 Suppl 9: 201 212.
39. Mostoslavsky R, Chua KF, Lombard DB, Pang WW, Fischer MR, et al. (2006) Genomic instability and aging-like phenotype in the absence of mammalian SIRT6. Cell 124: 315 329.
40. Lee SS, Kennedy S, Tolonen AC, Ruvkun G (2003) DAF-16 target genes that control C. elegans life-span and metabolism. Science 300: 644 647.
41. McElwee J, Bubb K, Thomas JH (2003) Transcriptional outputs of the Caenorhabditis elegans forkhead protein DAF-16. Aging Cell 2: 111 121.
42. Murphy CT, McCarroll SA, Bargmann CI, Fraser A, Kamath RS, et al. (2003) Genes that act downstream of DAF-16 to influence the lifespan of Caenorhabditis elegans. Nature 424: 277 283.
43. Mukhopadhyay A, Oh SW, Tissenbaum HA (2006) Worming pathways to and from DAF-16/FOXO. Exp Gerontol 41: 928 934.
44. Hsu AL, Murphy CT, Kenyon C (2003) Regulation of aging and age-related disease by DAF-16 and heat-shock factor. Science 300: 1142 1145.
45. Morley JF, Brignull HR, Weyers JJ, Morimoto RI (2002) The threshold for polyglutamine-expansion protein aggregation and cellular toxicity is dynamic and influenced by aging in Caenorhabditis elegans. Proc Natl Acad Sci U S A 99: 10417 10422.
46. Nasrin N, Ogg S, Cahill CM, Biggs W, Nui S, et al. (2000) DAF-16 recruits the CREB-binding protein coactivator complex to the insulin-like growth factor binding protein 1 promoter in HepG2 cells. Proc Natl Acad Sci U S A 97: 10412 10417.
47. Finnin MS, Donigian JR, Pavletich NP (2001) Structure of the histone deacetylase SIRT2. Nat Struct Biol 8: 621 625.
48. Senf SM, Sandesara PB, Reed SA, Judge AR (2011) p300 Acetyltransferase activity differentially regulates the localization and activity of the FOXO homologues in skeletal muscle. Am J Physiol Cell Physiol 300: C1490 1501.
49. Zhang M, Poplawski M, Yen K, Cheng H, Bloss E, et al. (2009) Role of CBP and SATB-1 in aging, dietary restriction, and insulin-like signaling. PLoS Biol 7: e1000245. doi:10.1371/journal.pbio.1000245
50. Eastburn DJ, Han M (2005) A gain-of-function allele of cbp-1, the Caenorhabditis elegans ortholog of the mammalian CBP/p300 gene, causes an increase in histone acetyltransferase activity and antagonism of activated Ras. Mol Cell Biol 25: 9427 9434.
51. Hunt-Newbury R, Viveiros R, Johnsen R, Mah A, Anastas D, et al. (2007) High-throughput in vivo analysis of gene expression in Caenorhabditis elegans. PLoS Biol 5: e237. doi:10.1371/journal.pbio.0050237
52. McKay SJ, Johnsen R, Khattra J, Asano J, Baillie DL, et al. (2003) Gene expression profiling of cells, tissues, and developmental stages of the nematode C. elegans. Cold Spring Harb Symp Quant Biol 68: 159 169.
53. Ko SI, Lee IS, Kim JY, Kim SM, Kim DW, et al. (2006) Regulation of histone acetyltransferase activity of p300 and PCAF by proto-oncogene protein DEK. FEBS Lett 580: 3217 3222.
54. Qiang L, Banks AS, Accili D (2010) Uncoupling of acetylation from phosphorylation regulates FoxO1 function independent of its subcellular localization. J Biol Chem 285: 27396 27401.
55. Brent MM, Anand R, Marmorstein R (2008) Structural basis for DNA recognition by FoxO1 and its regulation by posttranslational modification. Structure 16: 1407 1416.
56. Wang F, Chan CH, Chen K, Guan X, Lin HK, et al. (2012) Deacetylation of FOXO3 by SIRT1 or SIRT2 leads to Skp2-mediated FOXO3 ubiquitination and degradation. Oncogene 31: 1546 1557.
57. Pfister JA, Ma C, Morrison BE, D'Mello SR (2008) Opposing effects of sirtuins on neuronal survival: SIRT1-mediated neuroprotection is independent of its deacetylase activity. PLoS ONE 3: e4090. doi:10.1371/journal.pone.0004090
58. Pan PW, Feldman JL, Devries MK, Dong A, Edwards AM, et al. (2011) Structure and biochemical functions of SIRT6. J Biol Chem 286: 14575 14587.
59. Brenner S (1974) The genetics of Caenorhabditis elegans. Genetics 77: 71 94.
60. Mello CC, Kramer JM, Stinchcomb D, Ambros V (1991) Efficient gene transfer in C.elegans: extrachromosomal maintenance and integration of transforming sequences. EMBO J 10: 3959 3970.
61. Apfeld J, Kenyon C (1999) Regulation of lifespan by sensory perception in Caenorhabditis elegans. Nature 402: 804 809.
62. Kenyon C, Chang J, Gensch E, Rudner A, Tabtiang R (1993) A C. elegans mutant that lives twice as long as wild type. Nature 366: 461 464.
63. Ahn BH, Kim HS, Song S, Lee IH, Liu J, et al. (2008) A role for the mitochondrial deacetylase Sirt3 in regulating energy homeostasis. Proc Natl Acad Sci U S A 105: 14447 14452.